CSCI 6444 Big Data and Analytics

Class Project #2

Report

Text Analytics in R

Name: Tianyu Yang (G38878678)

Zhijun Xia (G23682615)

Date: 7/14/2019

Instructor: Stephen Kaisler

Term: 2019 Summer

1. Introduction

In project 2, our group tried to use what we learned from the class to do the text analytics in R language. The data set we used is Dream.txt written by Henri Bergson, which is a popular book and contains 54833 characters.

In this project, we will use several useful packages in R, such as textreuse, wordnet, zipfR, corpusTools, stringi, corpustools, quanteda and so on to generate 8 deliverables (a to h) from the text. The result will be displayed in the Result Section of this report.

Furthermore, we will write R functions to search through the documents to find three specific words or phrases, and display the result in the Word Search Section of this report.
Last but not the least, we will give out the conclusion of this project, which will give a summary of this project and show what we learn from this project and knowledge about data science in the Conclusion Section. All the R codes will be zipped together with this report as the name Group-2-Project-2.zip.

2. Result

2.a Lecture 4 Function Demonstration

Create a VCorpus:

```
> setwd("~/CSCI64442")
> getwd()
[1] "C:/Users/tangl/Documents/CSCI64442"
> SAT<-VCorpus(DirSource(".", ignore.case=TRUE, mode="text"))
> SAT
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1
```

Inspect the VCorpus:

```
> inspect(SAT)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 54850
```

VCorpus Structure:

```
> str(SAT)
List of 1
 $ pg20842.txt:List of 2
  ..$ content: chr [1:891] "ï»¿" "" "INTRODUCTION" "" ...
  ..$ meta   :List of 7
  .. ..$ author       : chr(0)
  .. ..$ datetimestamp: POSIXlt[1:1], format: "2019-07-14 21:54:26"
  .. ..$ description  : chr(0)
  .. ..$ heading      : chr(0)
  .. ..$ id           : chr "pg20842.txt"
  .. ..$ language     : chr "en"
  .. ..$ origin       : chr(0)
  .. ..- attr(*, "class")= chr "TextDocumentMeta"
  ..- attr(*, "class")= chr [1:2] "PlainTextDocument" "TextDocument"
 - attr(*, "class")= chr [1:2] "VCorpus" "Corpus"
```

Extract a document from SAT:

```
> test1<-SAT[[1]]
> test1
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 54850
> SATtdm <- TermDocumentMatrix(SAT)
> SATtdm
<<TermDocumentMatrix (terms: 2611, documents: 1)>>
Non-/sparse entries: 2611/0
Sparsity           : 0%
Maximal term length: 18
Weighting          : term frequency (tf)
```

Inspecting the TDM:

```
> inspect(SATtdm[1:10,1:1])
<<TermDocumentMatrix (terms: 10, documents: 1)>>
Non-/sparse entries: 10/0
Sparsity           : 0%
Maximal term length: 14
Weighting          : term frequency (tf)
Sample             :
               Docs
Terms           pg20842.txt
  "b----                  1
  "colored                1
  "consolation."          1
  "good                   1
  "ocular                 1
  "out!                   2
  "phosphenes,"           1
  "positively             1
  "preface                1
  "psychical              1
>
```

Document Term Frequency:

```
> test1tf <- termFreq(test1)
> test1tf
```

"b----	"colored	"consolation."	"good
1	1	1	1
"ocular	"out!	"phosphenes,"	"positively
1	2	1	1
"preface	"psychical	"railroad"	"railroad,"
1	1	1	1
"railroad."	"so,	"subjective"	"such
1	1	1	1
"the	"to	"trieste"	"trost,"
1	1	1	1
"tumult"	"tunnel."	"verzweiflung"	"wow-wows"
1	1	1	1
"you	(car,	(despair).	(such
1	1	1	1
(which	_â€lan	_as	_creative
1	1	1	1
disinterested	_fire_	_fire_.	_fire_;
1	2	1	1
_i.e._,	_incapable_	_indifferent_	_institut
1	1	1	1
laughter.	_matter	_rã´le_	_rã´les_
1	1	2	1
_revue	_tension_,	_the	_times_:
1	1	1	1
10,	1848.	1901.	1913,
1	1	2	2
1914.	26,	28,	30,
1	1	1	1
ability	able	abnormal	abolition
1	3	1	3
abound	about	above	abruptly
1	8	2	2
absent	absorbs	abstain	abstract
1	1	1	1
absurd	absurdity	absurdity.	abundance
3	1	1	1
abundance,	acceleration	acceptance	accepted
1	1	1	1
accepts	accessible	accidents,	accompanies
1	1	1	1
accomplish	accomplish,	accomplished	accordance
1	2	2	1
according	account	accounts	accumulated
3	1	1	1

...

```
> test1df <-as.data.frame(test1tf)
> test1df
```

...

```
"colored                1
"consolation."          1
"good                   1
"ocular                 1
"out!                   2
"phosphenes,"           1
"positively             1
"preface                1
"psychical              1
"railroad"              1
"railroad,"             1
"railroad."             1
"so,                    1
"subjective"            1
"such                   1
"the                    1
"to                     1
"trieste"               1
"trost,"                1
"tumult"                1
"tunnel."               1
"verzweiflung"          1
"wow-wows"              1
"you                    1
(car,                   1
(despair).              1
```

Corpus Management:

```
> SATlow <- tm_map(SAT, content_transformer(tolower))
> SATlow
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1
> removeNumPunct <-function(x) gsub("[^[:alpha:][:space:]]*", "", x)
> SATcl <- tm_map(SATlow,content_transformer(removeNumPunct))
> SATcltdm <- TermDocumentMatrix(SATcl)
> SATcltdm
<<TermDocumentMatrix (terms: 2053, documents: 1)>>
Non-/sparse entries: 2053/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : term frequency (tf)
```

```
> inspect(SATcltdm[1:10,1:1])
<<TermDocumentMatrix (terms: 10, documents: 1)>>
Non-/sparse entries: 10/0
Sparsity           : 0%
Maximal term length: 9
Weighting          : term frequency (tf)
Sample             :
           Docs
Terms       pg20842.txt
  ability             1
  able                3
  abnormal            1
  abolition           3
  abound              1
  about               8
  above               2
  abruptly            2
  absent              1
  absorbs             1
```

```
> myStopwords <- c(stopwords('english'))
> myStopwords
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"
 [10] "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"       "their"
 [28] "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"         "does"       "did"        "doing"      "would"
 [55] "should"     "could"      "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"
 [64] "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"    "isn't"
 [82] "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"
[100] "that's"     "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"
[109] "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"
[118] "until"      "while"      "of"         "at"         "by"         "for"        "with"       "about"      "against"
[127] "between"    "into"       "through"    "during"     "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"
[154] "how"        "all"        "any"        "both"       "each"       "few"        "more"       "most"       "other"
[163] "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"       "so"
[172] "than"       "too"        "very"
```

Removing stop words:

```
> SATstop <- tm_map(SATcl, removeWords, myStopwords)
> inspect(SATstop[1:1])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 38049
```

Remove sparse terms:

```
> SATtdm2 <- TermDocumentMatrix(SATstop,control = list(wordlengths = c(1,Inf)))
> SATtdm2
<<TermDocumentMatrix (terms: 1962, documents: 1)>>
Non-/sparse entries: 1962/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : term frequency (tf)
> SATnosparse<- removeSparseTerms(SATtdm2, 0.50)
> SATnosparse
<<TermDocumentMatrix (terms: 1962, documents: 1)>>
Non-/sparse entries: 1962/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : term frequency (tf)
> SATtdm3 <- TermDocumentMatrix(SATstop,control = list(wordlenths = c(1,Inf),weighting= weightBin))
> SATtdm3
<<TermDocumentMatrix (terms: 1962, documents: 1)>>
Non-/sparse entries: 1962/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : binary (bin)
> SATtdm4 <- TermDocumentMatrix(SATstop,control = list(wordlenths = c(1,Inf),weighting= weightTfIdf))
> SATtdm4
<<TermDocumentMatrix (terms: 1962, documents: 1)>>
Non-/sparse entries: 0/1962
Sparsity           : 100%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

Finding Frequent Words:

```
> freq.terms <-findFreqTerms(SATtdm2,lowfreq = 5)
> freq.terms
  [1] "action"        "almost"        "also"          "among"         "another"       "appear"        "appears"
  [8] "around"        "asleep"        "attention"     "awake"         "become"        "becomes"       "bergson"
 [15] "bergsons"      "besides"       "body"          "book"          "brain"         "can"           "capable"
 [22] "cases"         "certain"       "chance"        "closed"        "colors"        "come"          "common"
 [29] "complete"      "consciousness" "continue"      "day"           "depths"        "difference"    "different"
 [36] "difficult"     "dog"           "doubtless"     "dream"         "dreamer"       "dreaming"      "dreams"
 [43] "ear"           "effort"        "ego"           "either"        "enough"        "essential"     "etc"
 [50] "even"          "events"        "every"         "exactly"       "example"       "experience"    "explain"
 [57] "explanation"   "external"      "eyes"          "far"           "feel"          "feeling"       "field"
 [64] "finally"       "find"          "finds"         "fire"          "first"         "form"          "forms"
 [71] "general"       "give"          "given"         "great"         "hand"          "idea"          "image"
 [78] "images"        "impressions"   "internal"      "just"          "know"          "least"         "let"
 [85] "letters"       "life"          "light"         "like"          "little"        "live"          "living"
 [92] "made"          "make"          "making"        "man"           "many"          "material"      "matter"
 [99] "may"           "means"         "memories"      "memory"        "mental"        "might"         "mind"
[106] "moment"        "much"          "must"          "nature"        "necessary"     "never"         "nevertheless"
[113] "new"           "night"         "normal"        "nothing"       "now"           "objects"       "observations"
[120] "observer"      "often"         "one"           "order"         "others"        "part"          "past"
[127] "perceive"      "perceived"     "perception"    "perceptions"   "persons"       "place"         "play"
[134] "point"         "points"        "present"       "pressure"      "professor"     "psychical"     "question"
[141] "quite"         "räle"          "read"          "real"          "really"        "regard"        "remembrance"
[148] "requires"      "say"           "see"           "seen"          "sensation"     "sensations"    "sense"
[155] "senses"        "short"         "side"          "since"         "sleep"         "slumber"       "something"
[162] "sometimes"     "sounds"        "space"         "speak"         "spoken"        "spots"         "state"
[169] "still"         "study"         "subject"       "suppose"       "take"          "takes"         "tension"
[176] "theory"        "things"        "think"         "thought"       "thus"          "time"          "together"
[183] "toward"        "true"          "two"           "unconscious"   "upon"          "vague"         "visual"
[190] "waking"        "watch"         "way"           "well"          "white"         "whole"         "will"
[197] "without"       "word"          "words"         "written"
> findAssocs(SATtdm2,"states",0.25)
$states
numeric(0)

> freq.terms3 <-findFreqTerms(SATtdm3,lowfreq = 5)
> freq.terms3
character(0)
> freq.terms4 <-findFreqTerms(SATtdm4,lowfreq = 5)
> freq.terms4
character(0)
```

Term Frequency:

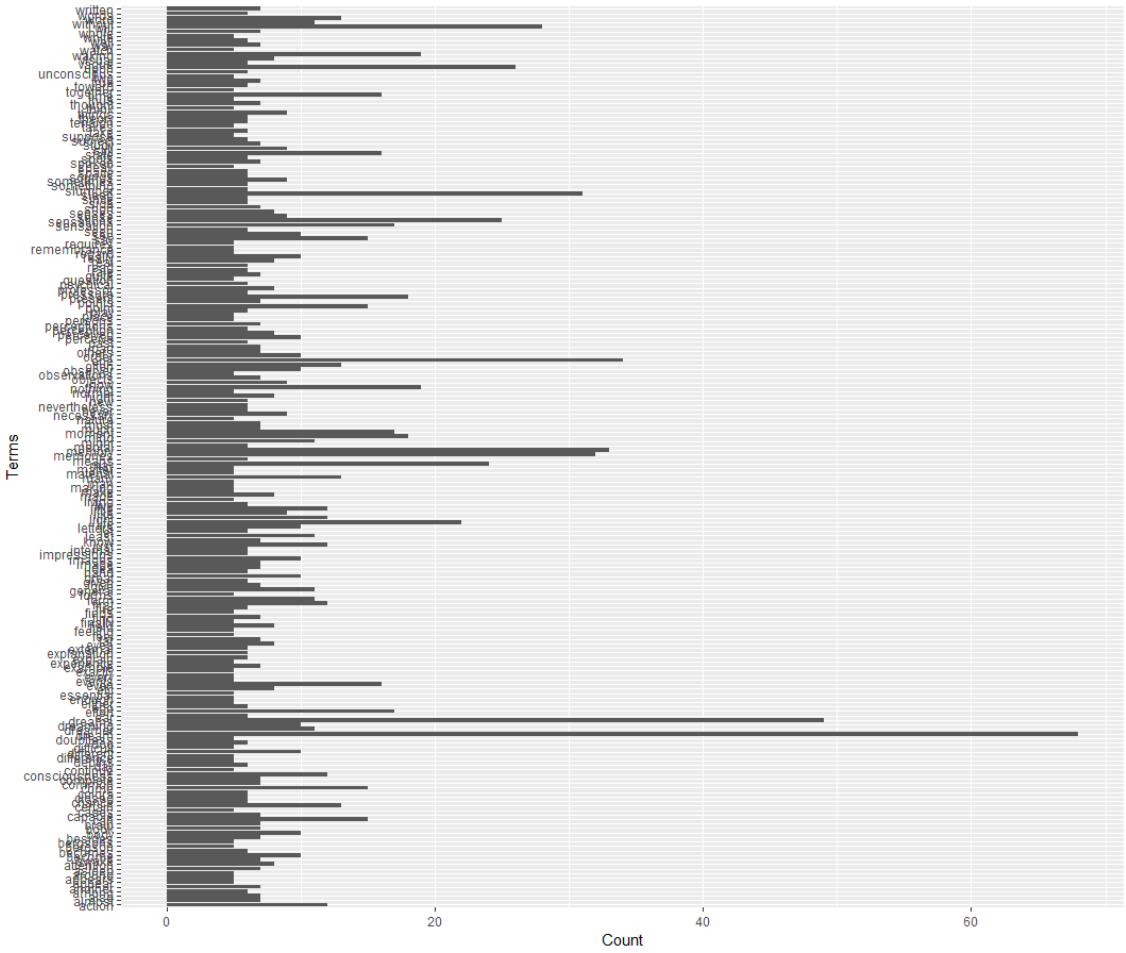
```
> term.freq <-rowSums(as.matrix(SATtdm2))
> term.freq <-subset(term.freq,term.freq>=5)
> df <-data.frame(term = names(term.freq), freq = term.freq)
> term.freq
       action        almost          also        among      another       appear      appears       around        asleep
           12             7             7            6            7            5            5            5             7
    attention         awake        become      becomes      bergson     bergsons      besides         body          book
            8             7            10            6            5            5            7           10             7
        brain           can       capable        cases      certain       chance       closed       colors         come
            7            15             7            5           13            6            6            6            15
       common      complete consciousness     continue          day       depths   difference    different    difficult
            7             7            12            5            6            5            5           10             5
          dog     doubtless         dream      dreamer     dreaming      dreams          ear       effort          ego
            6             5            68           11           10           49            6           17            6
       either        enough     essential          etc         even       events        every      exactly      example
            5             5             5            8           16            5            5            5            7
   experience       explain   explanation     external         eyes          far         feel      feeling        field
            5             6             6            6            8            7            5            5            8
      finally          find         finds         fire        first         form        forms      general         give
            5             7             5            6           12           11            5           11            7
        given         great          hand         idea        image       images  impressions     internal         just
            6            10             6            7            7           10            6            6           12
         know         least           let      letters         life        light         like       little         live
            7            11             6           10           22           12            9           12            6
       living          made          make       making          man         many     material       matter          may
            5             8             5            5            5           13            5            5           24
```

...

```
> df
                      term freq
action              action   12
almost              almost    7
also                  also    7
among                among    6
another            another    7
appear              appear    5
appears            appears    5
around              around    5
asleep              asleep    7
attention        attention    8
awake                awake    7
become              become   10
becomes            becomes    6
bergson            bergson    5
bergsons          bergsons    5
besides            besides    7
body                  body   10
book                  book    7
brain                brain    7
can                    can   15
capable            capable    7
cases                cases    5
certain            certain   13
chance              chance    6
closed              closed    6
colors              colors    6
come                  come   15
common              common    7
complete          complete    7
consciousness consciousness   12
continue          continue    5
day                    day    6
depths              depths    5
difference       difference    5
different         different   10
difficult         difficult    5
dog                    dog    6
doubtless         doubtless    5
...
```

Plot graph:

```
> ggplot(df,aes(x=term,y=freq))+geom_bar(stat="identity")+xlab("Terms")+ylab("Count")+coord_flip()
```



Clustering Terms:

```
> tdm2<- removeSparseTerms(SATtdm,sparse = 0.50)
> tdm2
<<TermDocumentMatrix (terms: 2611, documents: 1)>>
Non-/sparse entries: 2611/0
Sparsity           : 0%
Maximal term length: 18
Weighting          : term frequency (tf)
> disMatrix<-dist(scale(tdm2))
> disMatrix
                      "b----    "colored "consolation."       "good     "ocular       "out! "phosphenes," "positively
                   "preface  "psychical  "railroad" "railroad," "railroad."        "so, "subjective"        "such        "the
                       "to   "trieste"    "trost,"    "tumult"   "tunnel." "verzweiflung"  "wow-wows"        "you
                     (car,  (despair).       (such      (which      _â€lan          _as   _creative _disinterested_
                    _fire_     _fire_.     _fire_;     _i.e._, _incapable_ _indifferent_   _institut _laughter_.
                   _matter     _rã´le_    _rã´les_      _revue  _tension_,        _the    _times_:         10,       1848.
                     1901.       1913,       1914.         26,         28,         30,     ability        able    abnormal
                  abolition      abound       about       above    abruptly      absent     absorbs     abstain    abstract
                     absurd   absurdity  absurdity.   abundance  abundance, acceleration  acceptance    accepted     accepts
                 accessible  accidents, accompanies  accomplish accomplish, accomplished  accordance   according     account
                   accounts accumulated  acquainted   acquiring         act         act,      acting     acting,      action
                  action--in     action,     action.     active.    activity      actual     actually     acuity,       adapt
                 adaptation,     adapted         add     address    address,        adds  adherence,      adjust  adjustment
                 adjustment.     adjusts admission;"       admit     adopted   advantage  advantage, advantageous        afar
                      affect    affected   affection  affections affections,   affective       after       again     against
```

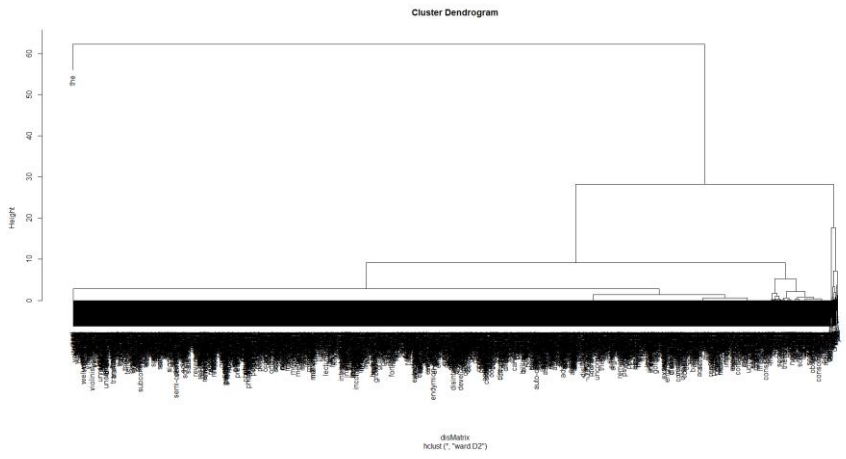
…

Finding Informative Words:

```
> test1 <-SATstop[[1]]
> inspect(test1)
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 38049

ï

introduction


  dawn  history mankind  engaged   study  dreaming
 wise man among  ancients  preâminently  interpreter
dreams  ability  interpret successfully  plausibly
quickest road  royal favor  joseph  daniel found
failure  give satisfaction   respect led  banishment
court  death   scholar laboriously translates  cuneiform tablet
dug    babylonian mound    lain buried  five
thousand years    chances     turn  either
astrological treatise   dream book   former  look upon
  indulgence   latter  pure contempt   know
 study   stars though undertaken  selfish reasons
pursued   spirit  charlatanry led  length  physical science
  study  dreams  proved  unprofitable   dreaming
   astrology grew astronomy   oneiromancy
grownnothing

  least  substantially true    beginning   present
century dream books   languages continued  sell  cheap
editions   interpreters  dreams made  decent    rate
comfortable living    poorer classes   psychologist
rarely paid attention  dreams except incidentally   study
imagery association   speed  thought  now  change  come
  spirit   times  subject   significance  dreams
 long ignored  suddenly become  matter  energetic study
fiery controversy  world

 cause   revival  interest   new point  view brought
forward  professor bergson   paper    made accessible
  englishreading public    idea   can explore
unconscious substratum   mentality  storehouse   memories
 means  dreams   memories    means inert
    life  purpose     strive  rise
consciousness whenever  get  chance even
semiconsciousness   dream  use professor bergsons striking
metaphor  memories  packed away  pressure like steam
boiler   dream   escape valve
```

...

Clustering Terms via Dendrogram:



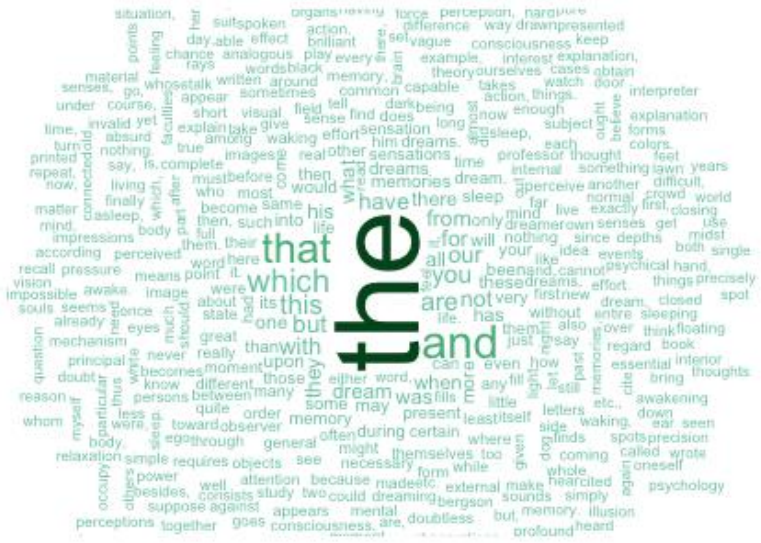
Word Cloud:

```
> m1<- as.matrix(tdm2)
> word.freq<-sort(rowSums(m1),decreasing = T)
> word.freq
```

the	and	that	which	this	are	you	but	our	with
715	221	190	126	91	80	74	69	65	64
have	not	for	from	they	dream	his	was	all	what
61	60	54	53	50	45	44	44	39	38
when	there	one	has	more	these	upon	memories	into	may
38	37	34	31	31	26	26	25	24	24
dreams	will	its	been	some	very	would	memory	than	during
23	23	22	21	21	20	20	19	19	18
only	present	then	your	sensations	sleep	those	dream.	other	come
18	18	18	18	17	17	17	16	16	15
even	life	can	dreams.	had	same	them	certain	many	such
15	15	14	14	14	14	14	13	13	13
their	any	here	just	most	nothing	often	sensation	were	effort
13	12	12	12	12	12	12	12	12	11
it,	it.	might	mind	moment	state	themselves	time	waking	become
11	11	11	11	11	11	11	11	11	10
before	great	him	least	little	order	really	then,	without	word
10	10	10	10	10	10	10	10	10	10
different	dreamer	dreams,	first	general	light	like	necessary	observer	out
9	9	9	9	9	9	9	9	9	9
point	see	should	them.	where	about	and,	because	between	does
9	9	9	9	9	8	8	8	8	8
form	how	itself	night	perceive	professor	real	say	sleep,	almost
8	8	8	8	8	8	8	8	8	7
being	capable	common	far	field	find	give	images	made	memory,
7	7	7	7	7	7	7	7	7	7
must	objects	part	quite	sometimes	through	visual	while	who	among
7	7	7	7	7	7	7	7	7	6
attention	becomes	cannot	complete	could	dreaming	external	given	idea	image
6	6	6	6	6	6	6	6	6	6
internal	let	letters	live	means	memories,	mental	much	never	new
6	6	6	6	6	6	6	6	6	6
now	own	perceived	read	sense	side	since	still	study	take
6	6	6	6	6	6	6	6	6	6
theory	toward	which,	action,	appears	around	bergson	body	brain	consciousness,
6	6	6	5	5	5	5	5	5	5
dog	doubtless	dream,	each	effort.	either	enough	events	every	exactly
5	5	5	5	5	5	5	5	5	5
example,	explain	eyes	feel	finds	know	life.	make	moment,	normal
5	5	5	5	5	5	5	5	5	5
observations	once	ourselves	over	persons	play	psychical	regard	requires	senses
5	5	5	5	5	5	5	5	5	5
short	sleep.	something	sounds	subject	takes	there,	things.	thought	too
5	5	5	5	5	5	5	5	5	5

...

```
> pal <- brewer.pal(9,"BuGn")
> pal <- pal[-(1:4)]
> wordcloud(words = names(word.freq),freq = word.freq,min.freq = 3,random.order = F,colors = pal)
```



Frequency Analysis:

```
> SATdtm <- DocumentTermMatrix(SATstop)
> freq<-colSums(as.matrix(SATdtm))
> SATdtm
<<DocumentTermMatrix (documents: 1, terms: 1962)>>
Non-/sparse entries: 1962/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : term frequency (tf)
> length(freq)
[1] 1962
>
> ord <-order(freq,decreasing = TRUE)
> freq[head(ord)]
   dream   dreams      one   memory memories    sleep
     68       49       34       33       32       31
> freq[tail(ord)]
     yale yellowish      yes     york    young     zeal
        1        1        1        1        1        1
>
> SATdtmr <-DocumentTermMatrix(SATstop,control = list(wordLengths=c(4,20)))
> SATdtmr
<<DocumentTermMatrix (documents: 1, terms: 1908)>>
Non-/sparse entries: 1908/0
Sparsity           : 0%
Maximal term length: 17
Weighting          : term frequency (tf)
>
> freqr<-colSums(as.matrix(SATdtmr))
> ordr <-order(freqr,decreasing = TRUE)
> freq[head(ordr)]
divination     domain     manner    mankind      sense    treated
         2          2          2          1          9          1
> freq[tail(ordr)]
vitality     wake   waking  wanders     want    watch
       2        2       19        1        1        5
```

Text Analysis:

```
> SAT1<-SAT$content[1]
> SAT1
[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 54850
```

```
> SAT1txt<-SAT1[[1]]$content
> SAT1txt
  [1] "ï»¿"
  [3] "INTRODUCTION"
  [5] ""
  [7] "The wise man among the ancients was preÃ«minently the interpreter of"
  [9] "quickest road to royal favor, as Joseph and Daniel found it to be;"
 [11] "court or death. When a scholar laboriously translates a cuneiform tablet"
 [13] "thousand years or more, the chances are that it will turn out either an"
 [15] "with some indulgence; if the latter with pure contempt. For we know that"
 [17] "pursued in the spirit of charlatanry, led at length to physical science,"
 [19] "them. Out of astrology grew astronomy. Out of oneiromancy has"
 [21] ""
 [23] "century. Dream books in all languages continued to sell in cheap"
 [25] "comfortable living out of the poorer classes. But the psychologist"
 [27] "imagery, association and the speed of thought. But now a change has come"
 [29] "so long ignored, has suddenly become a matter of energetic study and of"
 [31] ""
 [33] "forward by Professor Bergson in the paper which is here made accessible"
 [35] "unconscious substratum of our mentality, the storehouse of our memories,"
 [37] "as it were, a life and purpose of their own, and strive to rise into"
 [39] "semi-consciousness of a dream. To use Professor Bergson's striking"
 [41] "boiler and the dream is their escape valve."
 [43] "That this is more than a mere metaphor has been proved by Professor"
 [45] "inducing the patient to give expression to the secret anxieties and"
 [47] "clue to these disturbing thoughts is generally obtained in dreams or"
 [49] "dream always means something, but never what it appears to mean. It is"
 [51] "admit to consciousness, either because they are painful or because they"
 [53] "of consciousness to keep them back, but sometimes these unwelcome"
 [55] "this theory has developed the wildest extravagancies, and the voluminous"
 [57] "quite as absurd as the stuff which fills the twenty-five cent dream"
```

```
 [59] ""
 [61] ""
 [63] "Before the dawn of history mankind was engaged in the study of dreaming."
 [65] "dreams. The ability to interpret successfully or plausibly was the"
 [67] "failure to give satisfaction in this respect led to banishment from"
 [69] "dug up from a Babylonian mound where it has lain buried for five"
 [71] "astrological treatise or a dream book. If the former, we look upon it"
 [73] "the study of the stars, though undertaken for selfish reasons and"
 [75] "while the study of dreams has proved as unprofitable as the dreaming of"
 [77] "grown--nothing."
 [79] "That at least was substantially true up to the beginning of the present"
 [81] "editions and the interpreters of dreams made a decent or, at any rate, a"
 [83] "rarely paid attention to dreams except incidentally in his study of"
 [85] "over the spirit of the times. The subject of the significance of dreams,"
 [87] "fiery controversy the world over."
 [89] "The cause of this revival of interest is the new point of view brought"
 [91] "to the English-reading public. This is the idea that we can explore the"
 [93] "by means of dreams, for these memories are by no means inert, but have,"
 [95] "consciousness whenever they get a chance, even into the"
 [97] "metaphor, our memories are packed away under pressure like steam in a"
 [99] ""
[101] "Freud and others of the Vienna school, who cure cases of hysteria by"
[103] "emotions which, unknown to him, have been preying upon his mind. The"
[105] "similar states of relaxed consciousness. According to the Freudians a"
[107] "symbolic and expresses desires or fears which we refuse ordinarily to"
[109] "are repugnant to our moral nature. A watchman is stationed at the gate"
[111] "intruders slip past him in disguise. In the hands of fanatical Freudians"
[113] "literature of psycho-analysis contains much that seems to the layman"
[115] "book."
```

…

```
125  SAT1tokens<-tokens(SAT1txt)
126  SAT1tokens
128:1  (Top Level)                                                    R Script

Console  Terminal  Jobs
~/Desktop/project2/txt/

[13] "subsoil"     "of"

text884 :
 [1] "consciousness" ","             "that"          "will"          "be"
 [6] "the"           "principal"     "task"          "of"            "psychology"
[11] "in"            "the"

text885 :
 [1] "century"     "which"       "is"          "opening"     "."           "I"
 [7] "do"          "not"         "doubt"       "that"        "wonderful"   "discoveries"

text886 :
 [1] "await"     "it"        "there"     ","         "as"        "important" "perhaps"
 [8] "as"        "have"      "been"      "in"        "the"       "preceding"

text887 :
 [1] "centuries"   "the"         "discoveries" "of"          "the"         "physical"
 [7] "and"         "natural"     "sciences"    "."           "That"        "at"

text888 :
 [1] "least"   "is"      "the"     "promise" "which"   "I"       "make"    "for"     "it"
[10] ","       "that"    "is"      "the"     "wish"    "that"    "in"

text889 :
[1] "closing" "I"       "have"    "for"     "it"      "."

text890 :
character(0)

text891 :
character(0)
```

Text Analysis: Sentiment Analysis:

```
> SAT1Sent<- syuzhet::get_nrc_sentiment(SAT1txt)
> SAT1Sent
    anger anticipation disgust fear joy sadness surprise trust negative positive
1       0            0       0    0   0       0        0     0        0        0
2       0            0       0    0   0       0        0     0        0        0
3       0            0       0    0   0       0        0     0        0        0
4       0            0       0    0   0       0        0     0        0        0
5       0            0       0    0   0       0        0     0        0        0
6       0            2       0    0   2       0        1     2        0        3
7       0            0       0    0   0       0        0     0        0        1
8       0            0       0    0   0       0        0     0        0        1
9       0            0       0    0   1       0        0     1        0        1
10      1            1       2    1   1       2        0     1        2        1
11      2            2       1    2   0       1        1     0        1        1
12      0            0       0    1   0       1        0     0        1        0
13      0            0       0    0   0       0        0     0        0        0
14      0            0       0    0   0       0        0     0        0        0
15      1            0       1    1   0       0        0     0        1        0
16      1            0       1    0   0       0        0     0        1        1
17      0            0       0    0   0       0        0     0        0        1
18      0            0       0    0   0       0        0     0        1        1
19      0            0       0    0   0       0        0     0        0        0
20      0            0       0    0   0       0        0     0        0        0
21      0            0       0    0   0       0        0     0        0        0
22      0            1       0    0   1       0        1     1        0        1
23      0            0       0    0   0       0        0     0        1        0
24      0            0       0    0   0       0        0     0        0        1
25      0            0       0    0   0       0        0     0        0        0
```

…

```
> SAT1sdt<-rowSums(SAT1Sent)
> SAT1sdt
  [1]  0  0  0  0  0 10  1  1  3 12 11  3  0  0  4  4  1  2  0  0  0  5  1  1  0  2  3  2  4  1  0  5  4  4  1  1  1  2  4  1  4  0  2  3  3  3  1  1  0  2  6  6  3  2  2
 [56]  0  1  0  0  2  4  0  2  0  2  5 13  4  0  5  0  4  0 12  1 14  6  1  1 15  5  0  4  5  5  3  1  2  2  2  1  3  1  4  4  0  5  2  1  0  6  2  1  1  7  0  8  1  0  0
[111]  5  6  1  1  0  4  2  1  1  3  0  1  7  2  1  6  4  0  3  4  0  9  1  0  1  2  4  0  0  0  0  3  3  1  0  0  1  3  0  0  3  4  0  5  2  2  0  1  1  2  1  1  1  1  1
[166]  1  3  5  0  0  2  4  1  0  3  1  0  2  1  1  3  1  0  2  0  1  3  0  0  8  2  1  8  3  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  1  1  0  1  0  2  0  2  0  1
[221]  0  0  1  0  0  0  2  0  3  2  4  6  3  2  5  3  2  3  0  0  0  2  1  0  7  0  2  1  0  1  2  0  0  0  4  2  2  2  0  0  4  0  0  0  5  4  2  1  6  4  4  1  3  0  0
[276]  0  3  0  0  2  1  0  2  1  0  1  1  2  1  2  6  3  2  1  1  3  7  1  3  1  0  0  0  0  4  1  0  0  0  2  0  3  0  2  0  0  4  3  1  5  0  0  3  0  3  0  0  2  0  1
[331]  1  1  0  5  0  2  0  2  0  0  0  3  1  1  1  1  0  0  5  0  0  0  2 12  1  1  5  2  1  7  0  2  6  5  1  0  1  3  3  1  1  4  0  0  2  3  1  5  2  6 12 20  0  0  2
[386]  4  2  0  7  0  0  0  2  0  0  1  1  5  3  0  1  3  0  2  3  3  0  0  1  9  3  4  6  0  4  0  0  0  0  0  0  3  1  5  0  1  3  1  0  1  0  0  0  0  1  3  0  3  4  1
[441]  0  1  0  4  3  2  1  4  6  2  0  2  0  2  0  2  2  1  1  2  0  2  7  7  3  0  0  0  0  3  0  0  0  4  2  7  6  2  1  0  2  0  0  3  0  4  0  3  0  0  1  5  7  1  0
[496]  0  6  0  1  4  2  7  1  6  2  0  1  0  1  1  1  0  3  2  1  1  6  3  0  4  1  3  0  0  1  2  5  1  2  1  0  2  4  5  3  1  3  0  0  2  0  0  1  0  2  0  0  0  0  0
[551]  2  2  2  2  0  1  3  1  1  2  1  0  0  7  2  4  0  0  0  4  3  2  0  1  2  7  0  0  3  1  4  2  3  3  0  2  2  0  6  2  0  0  0  2  0  2  4  0  3  0  3  0  2  2  1
[606]  1  1  0  0  3  5  2  0  1  1  7  2  1  0  2  1  1  0  3  0  0  0  2  2  8  2  3  2  2  4  2  6  2  7  0  3  0  4  6  6  1  0  0  2  1  4  0  1  0  3  0  0  0  0  0
[661]  0  0  0  4  0  0  5  2  2  9  0  4  2  5  2  5  3  3  3  2  2  0  1  0  0  0  1  7  3  2  0  2  3  2  1  3  0  2  3  0 14  1  0  2  0  2  0  0  0  1  0  0  0  0  0
[716]  5  1  1  0  0  5  4  0  1  0  1  0  6  0  0  5  0  1 10  2  1  2  1  2  0  0  0  3  0  1  1  7  3  3  1  0  0  0  3  6  0  2  2  2  0  2  0  0  2  1  0  0  0  4  1
[771]  0  4  1  4  1  7  2  0  1  1  3  0  0  1  0  2  2  7  5  5  0  2  2  1  0  5  4  0  0  1  2  2  1  1  6  1  0  0  3  5  2  1  2  3  0  3  1  0  3  2  2  9  2  0  1
[826]  0  2  0  8  1  1  3  0  0  0  3  2  0  0  1  5  3  0  1  0  4  0  3 13  2  4  5  4  5  1  1  0  3  3  0  1  0  3  3 10  0  3  6  0 13  0  0  2  3  2  4  6  0  0  1
[881]  1  4  6  4  8  3  0  3  0  0  0
```

```
> SAT1rdt<-rowSums(SAT1Sent)
> SAT1rdt
  [1]  0  0  0  0  0 10  1  1  3 12 11  3  0  0  4  4  1  2  0  0  0  5  1  1  0  2  3  2  4  1  0  5  4  4  1  1  1  2  4  1  4  0  2  3  3  3  1  1  0  2  6  6  3  2  2
 [56]  0  1  0  0  2  4  0  2  0  2  5 13  4  0  5  0  4  0 12  1 14  6  1  1 15  5  0  4  5  5  3  1  2  2  2  1  3  1  4  4  0  5  2  1  0  6  2  1  1  7  0  8  1  0  0
[111]  5  6  1  1  0  4  2  1  1  3  0  1  7  2  1  6  4  0  3  4  0  9  1  0  1  2  4  0  0  0  0  3  3  1  0  0  1  3  0  0  3  4  0  5  2  2  0  1  1  2  1  1  1  1  1
[166]  1  3  5  0  0  2  4  1  0  3  1  0  2  1  1  3  1  0  2  0  1  3  0  0  8  2  1  8  3  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  1  1  0  1  0  2  0  2  0  1
[221]  0  0  1  0  0  0  2  0  3  2  4  6  3  2  5  3  2  3  0  0  0  2  1  0  7  0  2  1  0  1  2  0  0  0  4  2  2  2  0  0  4  0  0  0  5  4  2  1  6  4  4  1  3  0  0
[276]  0  3  0  0  2  1  0  2  1  0  1  1  2  1  2  6  3  2  1  1  3  7  1  3  1  0  0  0  0  4  1  0  0  0  2  0  3  0  2  0  0  4  3  1  5  0  0  3  0  3  0  0  2  0  1
[331]  1  1  0  5  0  2  0  2  0  0  0  3  1  1  1  1  0  0  5  0  0  0  2 12  1  1  5  2  1  7  0  2  6  5  1  0  1  3  3  1  1  4  0  0  2  3  1  5  2  6 12 20  0  0  2
[386]  4  2  0  7  0  0  0  2  0  0  1  1  5  3  0  1  3  0  2  3  3  0  0  1  9  3  4  6  0  4  0  0  0  0  0  0  3  1  5  0  1  3  1  0  1  0  0  0  0  1  3  0  3  4  1
[441]  0  1  0  4  3  2  1  4  6  2  0  2  0  2  0  2  2  1  1  2  0  2  7  7  3  0  0  0  0  3  0  0  0  4  2  7  6  2  1  0  2  0  0  3  0  4  0  3  0  0  1  5  7  1  0
[496]  0  6  0  1  4  2  7  1  6  2  0  1  0  1  1  1  0  3  2  1  1  6  3  0  4  1  3  0  0  1  2  5  1  2  1  0  2  4  5  3  1  3  0  0  2  0  0  1  0  2  0  0  0  0  0
[551]  2  2  2  2  0  1  3  1  1  2  1  0  0  7  2  4  0  0  0  4  3  2  0  1  2  7  0  0  3  1  4  2  3  3  0  2  2  0  6  2  0  0  0  2  0  2  4  0  3  0  3  0  2  2  1
[606]  1  1  0  0  3  5  2  0  1  1  7  2  1  0  2  1  1  0  3  0  0  0  2  2  8  2  3  2  2  4  2  6  2  7  0  3  0  4  6  6  1  0  0  2  1  4  0  1  0  3  0  0  0  0  0
[661]  0  0  0  4  0  0  5  2  2  9  0  4  2  5  2  5  3  3  3  2  2  0  1  0  0  0  1  7  3  2  0  2  3  2  1  3  0  2  3  0 14  1  0  2  0  2  0  0  0  1  0  0  0  0  0
[716]  5  1  1  0  0  5  4  0  1  0  1  0  6  0  0  5  0  1 10  2  1  2  1  2  0  0  0  3  0  1  1  7  3  3  1  0  0  0  3  6  0  2  2  2  0  2  0  0  2  1  0  0  0  4  1
[771]  0  4  1  4  1  7  2  0  1  1  3  0  0  1  0  2  2  7  5  5  0  2  2  1  0  5  4  0  0  1  2  2  1  1  6  1  0  0  3  5  2  1  2  3  0  3  1  0  3  2  2  9  2  0  1
[826]  0  2  0  8  1  1  3  0  0  0  3  2  0  0  1  5  3  0  1  0  4  0  3 13  2  4  5  4  5  1  1  0  3  3  0  1  0  3  3 10  0  3  6  0 13  0  0  2  3  2  4  6  0  0  1
[881]  1  4  6  4  8  3  0  3  0  0  0
```

```
> SAT1CDT<-colSums(SAT1Sent)
> SAT1CDT
       anger anticipation      disgust         fear          joy      sadness     surprise        trust     negative 
          89          193           57          130          113          106           97          241          252 
    positive 
         461 
```

Text Weighting:

```
> SATdfm <-dfm(SAT1txt)
> SATfrq <-docfreq(SATdfm)
> SATfrq
           ï             »             ¿  introduction        before           the 
           1             1             1             1            10           491 
        dawn            of       history       mankind           was       engaged 
           1           352             2             1            41             1 
          in         study      dreaming             .          wise           man 
         193             7            10           420             1             5 
       among      ancients          preã             «      minently   interpreter 
           6             1             1             1             1             3 
      dreams       ability            to     interpret  successfully            or 
          48             1           234             1             1            36 
   plausibly      quickest          road         royal         favor             , 
           1             1             1             1             1           456 
          as        joseph           and        daniel         found            it 
          62             1           214             1             2           152 
          be             ;       failure          give  satisfaction          this 
          51            35             1             7             1            90 
     respect           led    banishment          from         court         death 
           2             2             1            52             1             2 
        when             a       scholar   laboriously    translates     cuneiform 
          36           179             1             1             1             1 
      tablet           dug            up     babylonian         mound         where 
           1             1            10             1             1             9 
         has          lain        buried           for          five      thousand 
          30             1             1            53             1             2 
       years          more       chances           are          that          will 
           3            33             1            82           174            27 
        turn           out        either            an   astrological      treatise 
           4            13             5            34             1             1 
```

…

```
> SATweights2 <-dfm_weight(SATdfm,scheme = "prop")
> str(SATweights2)
Formal class 'dfm' [package "quanteda"] with 15 slots
  ..@ settings    : list()
  ..@ weightTf    :List of 3
  .. ..$ scheme: chr "prop"
  .. ..$ base  : NULL
  .. ..$ K     : NULL
  ..@ weightDf    :List of 5
  .. ..$ scheme   : chr "unary"
  .. ..$ base     : NULL
  .. ..$ c        : NULL
  .. ..$ smoothing: NULL
  .. ..$ threshold: NULL
  ..@ smooth      : num 0
  ..@ ngrams      : int 1
  ..@ skip        : int 0
  ..@ concatenator: chr "_"
  ..@ version     : int [1:3] 1 5 0
  ..@ docvars     :'data.frame':        891 obs. of  0 variables
  ..@ i           : int [1:10034] 0 0 0 2 5 133 172 307 355 436 ...
  ..@ p           : int [1:2131] 0 1 2 3 4 14 505 506 858 860 ...
  ..@ Dim         : int [1:2] 891 2130
  ..@ Dimnames    :List of 2
  .. ..$ docs    : chr [1:891] "text1" "text2" "text3" "text4" ...
  .. ..$ features: chr [1:2130] "i" "»" "¿" "introduction" ...
  ..@ x           : Named num [1:10034] 0.3333 0.3333 0.3333 1 0.0714 ...
  .. ..- attr(*, "names")= chr [1:10034] "text1" "text1" "text1" "text3" ...
  ..@ factors     : list()
> SATtfidf<-dfm_tfidf(SATdfm,scheme_tf = "count",scheme_df = "inverse")
> SATtfidf@i
   [1]   0   0   0   2   5 133 172 307 355 436 475 581 697 698   5   6   7  12  13  14  15  16  17  21  23  24  26  27  29  31  32  33  34  37  40  43  44  45  47  51  53
  [42]  54  55  56  59  60  61  65  66  69  71  72  73  74  76  77  78  84  85  88  89  90  91  92  93  96  97 100 101 102 103 104 105 106 107 114 116 117 122 124 126 128
  [83] 132 133 134 135 136 140 141 142 144 145 146 147 149 150 151 152 154 156 157 160 161 162 163 166 170 171 172 173 174 175 178 179 181 182 185 189 190 191 192 207 212
 [124] 213 225 229 233 236 238 240 241 242 243 244 245 246 249 250 252 257 258 259 261 262 263 264 266 267 270 272 274 276 279 280 281 282 285 286 287 288 289 291 292 293
 [165] 296 298 302 304 305 307 309 310 313 314 318 319 320 321 323 324 325 331 335 340 341 342 343 345 347 348 349 350 353 355 356 357 358 361 362 363 364 365 367 368 369
 [206] 371 372 375 382 383 385 386 388 389 390 392 393 394 395 402 403 407 408 410 413 414 416 427 430 431 435 438 439 440 441 443 445 446 447 448 449 450 454 456 458 461
 [247] 462 464 465 468 470 476 478 479 480 482 483 484 485 486 487 492 493 494 495 496 497 498 501 504 506 507 509 510 511 512 516 518 519 520 521 522 523 525 526 529 530
 [288] 531 533 534 536 537 539 541 542 543 545 546 547 548 549 550 551 552 553 554 555 556 557 559 563 564 566 568 569 570 573 577 578 580 582 583 584 585 586 587 588 589
 [329] 591 592 596 597 598 599 601 602 605 609 611 612 614 617 618 620 623 628 629 630 631 633 634 635 636 637 638 639 640 642 643 644 647 648 649 650 652 659 660 661 664
 [370] 665 666 667 668 669 673 674 675 680 682 683 687 689 690 693 695 697 698 699 701 703 704 707 711 718 719 720 721 722 723 724 725 726 731 735 737 738 741 745 746 747
```

…

2.b Longest sentences

(1) To obtain the longest sentences in the text, firstly we need to get the vectors contains every sentence in the text, shown as below.

```{r}
# Values preparation

# Chunk txt into sentences
chunk_into_sentences <- function(text) {
  break_points <- c(1, as.numeric(gregexpr('[[:alnum:] ][.!?]', text)[[1]]) + 1)
  sentences <- NULL
  for(i in 1:length(break_points)) {
    res <- substr(text, break_points[i], break_points[i+1])
    if(i>1) {
      sentences[i] <- sub('. ', '', res)
      # Remove useless punctuation
      sentences[i] <- gsub("\n", "", sentences[i]) #new line
      sentences[i] <- gsub("\", \"", " ", sentences[i]) #line ender
      sentences[i] <- gsub(".\"\"", "", sentences[i]) #header
      sentences[i] <- gsub("  ", "", sentences[i]) #space
      sentences[i] <- gsub("_", "", sentences[i]) #underline
      sentences[i] <- gsub("--", " ", sentences[i]) #connector
      } else { sentences[i] <- res }
    }
  sentences <- sentences[sentences=!is.na(sentences)]

  # Remove chapter name
  sentences <-removeWords(sentences,"Prpr")
  sentences <-removeWords(sentences,"INTRODUCTION")
  sentences <-removeWords(sentences,"DREAMS")
  return(sentences)
}

mycorpus <- VCorpus(VectorSource(SATtxt))

corpus_frame <- data.frame(text=unlist(sapply(mycorpus, `[`, "content")), stringsAsFactors=F)
sentences <-chunk_into_sentences(corpus_frame)
sentences
```

```
  [3] "The ability to interpret successfully or plausibly was the quickest road to royal favor, as Joseph and
Daniel found it to be; failure to give satisfaction in this respect led to banishment from court or death."
  [4] "When a scholar laboriously translates a cuneiform tablet dug up from a Babylonian mound where it has lain
buried for five thousand years or more, the chances are that it will turn out either an astrological treatise or a
dream book."
  [5] "If the former, we look upon it with some indulgence; if the latter with pure contempt."
  [6] "For we know that the study of the stars, though undertaken for selfish reasons and pursued in the spirit of
charlatanry, led at length to physical science, while the study of dreams has proved as unprofitable as the
dreaming of them."
  [7] "Out of astrology grew astronomy."
  [8] "Out of oneiromancy has grown nothing."
  [9] " That at least was substantially true up to the beginning of the present century."
 [10] "Dream books in all languages continued to sell in cheap editions and the interpreters of dreams made a
decent or, at any rate, a comfortable living out of the poorer classes."
 [11] "But the psychologist rarely paid attention to dreams except incidentally in his study of imagery,
association and the speed of thought."
 [12] "But now a change has come over the spirit of the times."
 [13] "The subject of the significance of dreams, so long ignored, has suddenly become a matter of energetic study
and of fiery controversy the world over."
 [14] " The cause of this revival of interest is the new point of view brought forward by Professor Bergson in the
paper which is here made accessible to the English-reading public."
 [15] "This is the idea that we can explore the unconscious substratum of our mentality, the storehouse of our
memories, by means of dreams, for these memories are by no means inert, but have, as it were, a life and purpose of
their own, and strive to rise into consciousness whenever they get a chance, even into the semi-consciousness of a
dream."
 [16] "To use Professor Bergson's striking metaphor, our memories are packed away under pressure like steam in a
boiler and the dream is their escape valve."
 [17] " That this is more than a mere metaphor has been proved by Professor Freud and others of the Vienna school,
who cure cases of hysteria by inducing the patient to give expression to the secret anxieties and emotions which,
unknown to him, have been preying upon his mind."
 [18] "The clue to these disturbing thoughts is generally obtained in dreams or similar states of relaxed
consciousness."
 [19] "According to the Freudians a dream always means something, but never what it appears to mean."
 [20] "It is symbolic and expresses desires or fears which we refuse ordinarily to admit to consciousness, either
because they are painful or because they are repugnant to our moral nature."
 [21] "A watchman is stationed at the gate of consciousness to keep them back, but sometimes these unwelcome
intruders slip past him in disguise."
 [22] "In the hands of fanatical Freudians this theory has developed the wildest extravagancies, and the voluminous
literature of psycho-analysis contains much that seems to the layman quite as absurd as the stuff which fills the
twenty-five cent dream book."
 [23] " It is impossible to believe that the subconsciousness of every one of us contains nothing but the foul and
monstrous specimens which they dredge up from the mental depths of their neuropathic patients and exhibit with such
```

(2) Write two R functions to realize the function. The first one is to obtain the index of the sentence that contains the most words. The second one is to obtain n longest sentences in the text. It is first to get the longest sentences and then remove it from the text. Repeatedly n times, then we can get n longest sentences. When n = 10, it is shown as below:

```{r}
# Project_2 1.(b)

# Find longest sentences
Find_longest_sentence <- function(text){
  result<-NULL
  wordCount <-NULL
  for(i in 1:length(text)){
    wordCount[i] = count_words(text[i])
  }
  max <-max(wordCount)
  for(i in 1:length(wordCount)){
    if(wordCount[i] == max)
      index <-i
  }
  return(index)
}

# Find 10 longest sentences
Find_sentences <- function(text,n){
  result <- NULL
  for(i in 1:n){
    result <-append(result,text[Find_longest_sentence(text)])
    text <-text[-Find_longest_sentence(text)]
  }
  return(result)
}

# Display result
Find_sentences(sentences,10)
```

```
 [1] "We need not be astonished, then, that philosophers like Schopenhauer have seen in the dream a reverberation,
in the heart of consciousness, of perturbations emanating from the sympathetic nervous system; and that
psychologists like Schemer have attributed to each of our organs the power of provoking a well-determined kind of
dream which represents it, as it were, symbolically; and finally that physicians like Artigues have written
treatises on the semeiological value of dreams, that is to say, the method of making use of dreams for the
diagnosis of certain maladies."
 [2] "When in the Palace of Night scene of his fairy play, the redoubtable Tyltyl unlocks the cage where are
confined the nightmares and all other evil imaginings, he shuts the door in time to keep them in and then opens
another revealing a lovely garden full of blue birds, which, though they fade and die when brought into the light
of common day, yet encourage him to continue his search for the Blue Bird that never fades, but lives
everlastingly."
 [3] " The subject which I have to discuss here is so complex, it raises so many questions of all kinds, difficult,
obscure, some psychological, others physiological and metaphysical; in order to be treated in a complete manner it
requires such a long development and we have so little space, that I shall ask your permission to dispense with all
preamble, to set aside unessentials, and to go at once to the heart of the question."
 [4] "It is sufficient for me to say, in order to answer the question which I have propounded, that the formative
power of the materials furnished to the dream by the different senses, the power which converts into precise,
determined objects the vague and indistinct sensations that the dreamer receives from his eyes, his ears, and the
whole surface and interior of his body, is the memory."
 [5] " Professor Bergson's theory of dreaming here set forth in untechnical language, fits into a particular niche
in his general system of philosophy as well as does his little book on Laughter. With the main features of his
philosophy the English-reading public is better acquainted than with any other contemporary system, for his books
have sold even more rapidly here than in France."
 [6] "When the mind creates, I would say when it is capable of giving the effort of organization and synthesis
which is necessary to triumph over a certain difficulty, to solve a problem, to produce a living work of the
imagination, we are not really asleep, or at least that part of ourselves which labors is not the same as that
which sleeps."
 [7] "This is the idea that we can explore the unconscious substratum of our mentality, the storehouse of our
memories, by means of dreams, for these memories are by no means inert, but have, as it were, a life and purpose of
their own, and strive to rise into consciousness whenever they get a chance, even into the semi-consciousness of a
dream."
 [8] "It would not beat all surprising if perceptions of the organs of our senses,useful perceptions, were the
result of a selection or of acanalization worked by the organs of our senses in the interest ofour action, but that
there should yet be around those perceptions afringe of vague perceptions, capable of becoming more distinct
inextraordinary, abnormal cases."
 [9] "In this address, also, was brought into consideration for the first time the idea that the self may go
through different degrees of tension a theory referred to in his Matter and Memory.Its chief interest for the
general reader will, however, lie in the explanation it gives him of the cause of some of his familiar dreams."
[10] "It depends especially upon three points, which are: the incoherence of dreams, the abolition of the sense of
duration that often appears to be manifested in dreams, and, finally, the order in which the memories present
themselves to the dreamer, contending for the sensations present where they are to be embodied."
```

2.c Dendrogram and WordCloud for each paragraph

To obtain the dendrogram and wordCloud for each paragraph, we write two functions to create dendrogram and wordCloud.

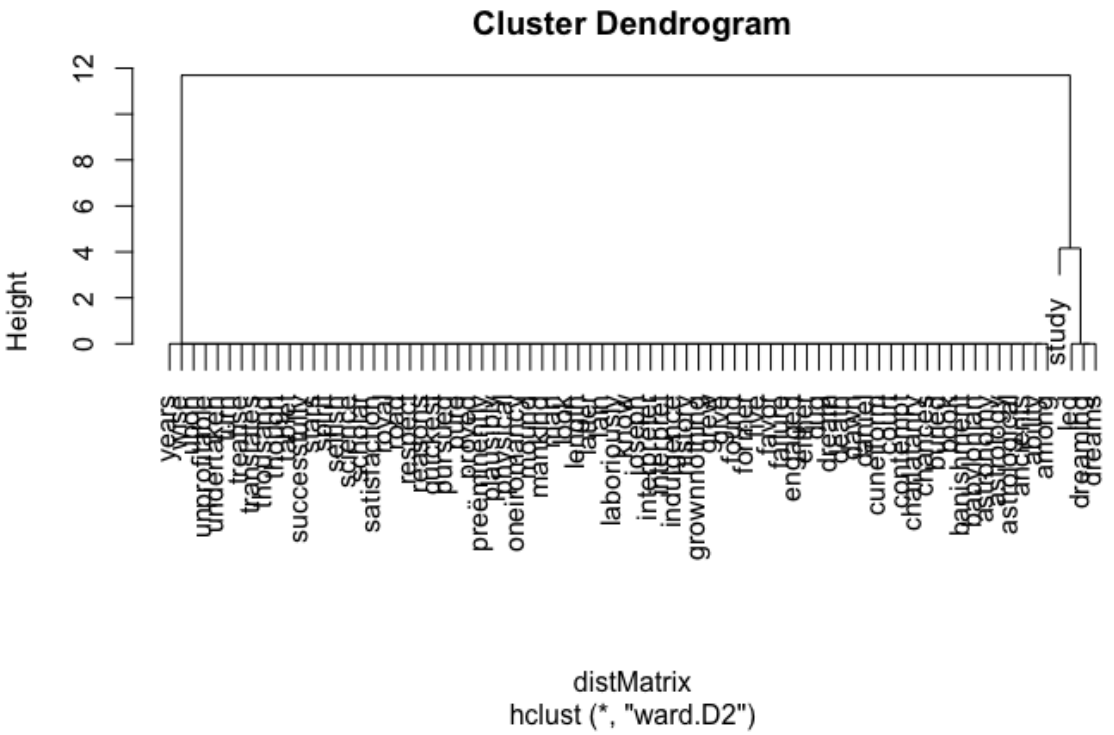
```
Create_dendrogram <-function(corpus){
  # Corpus Management
  SATlow <- tm_map(corpus, content_transformer(tolower))
  removeNumPunct <-function(x) gsub("[^[:alpha:][:space:]]*", "", x)
  SATcl <- tm_map(SATlow,content_transformer(removeNumPunct))
  myStopwords <- c(stopwords('english'))

  # Removing stop words
  SATstop <- tm_map(SATcl, removeWords, myStopwords)

  # Remove sparse terms
  SATtdm2 <- TermDocumentMatrix(SATstop,control = list(wordlengths = c(1,Inf)))

  # Clustering Terms
  distMatrix <-dist(scale(SATtdm2))

  ## Clustering Terms via Dendrogram
  fit <-hclust(distMatrix,method = "ward.D2")
  plot(fit)
}
Create_WordCloud <-function(corpus){
  # Corpus Management
  SATlow <- tm_map(SAT[2], content_transformer(tolower))
  removeNumPunct <-function(x) gsub("[^[:alpha:][:space:]]*", "", x)
  SATcl <- tm_map(SATlow,content_transformer(removeNumPunct))
  myStopwords <- c(stopwords('english'))

  # Removing stop words
  SATstop <- tm_map(SATcl, removeWords, myStopwords)

  # Remove sparse terms
  SATtdm2 <- TermDocumentMatrix(SATstop,control = list(wordlengths = c(1,Inf)))

  ## Word Cloud
  m1<- as.matrix(SATtdm2)
  word.freq <-sort(rowSums(m1),decreasing = T)
  pal <- brewer.pal(9,"BuGn")
  pal <- pal[-(1:4)]
  wordcloud(words = names(word.freq),freq = word.freq,min.freq = 3,random.order = F,colors = pal)
}
```

Before plot the figure, we need to finish the data cleaning. Firstly, we remove numbers and punctuation in the text. Then we clear all stopwords in ‘English’. Finally we list all the word tokens in lower level.

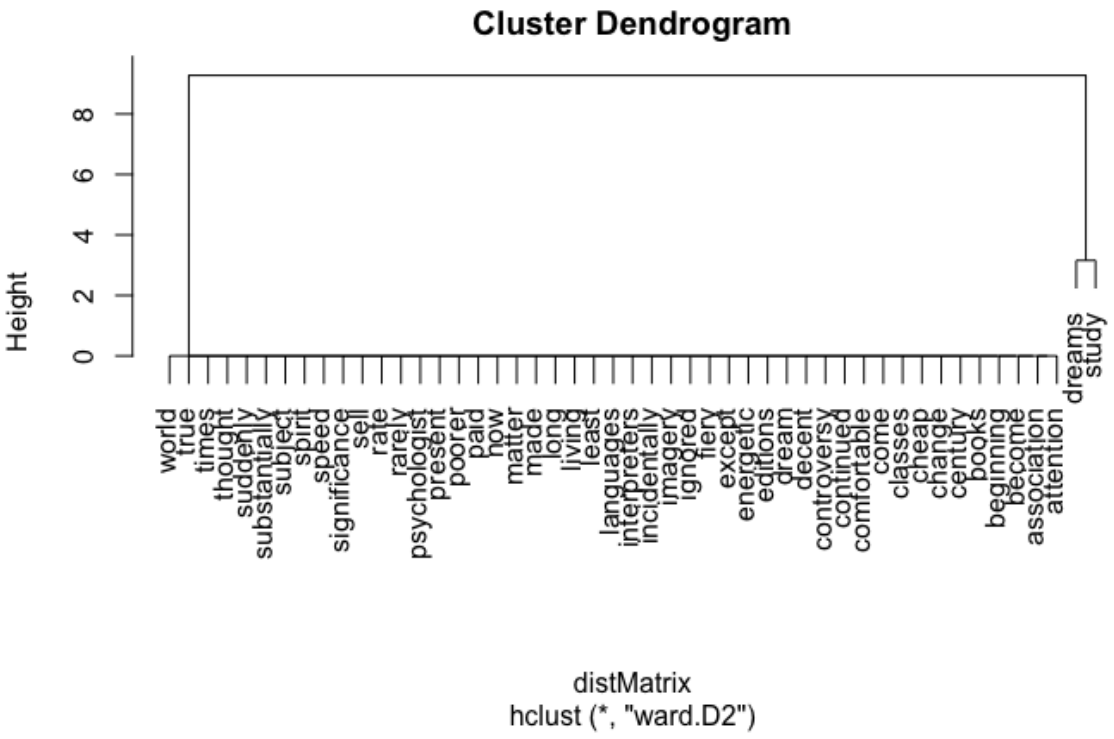
From 2.d, we separate the text into paragraphs. Therefore, we use for-loop to display the dendrogram and wordCloud for all the paragraphs in the text. Because of the limitation of the space of the report, we only display five of each as examples. For there is a limitation for the length of the paragraph, therefore we choose 1-4, 7 as examples.

(1) Dendrogram:

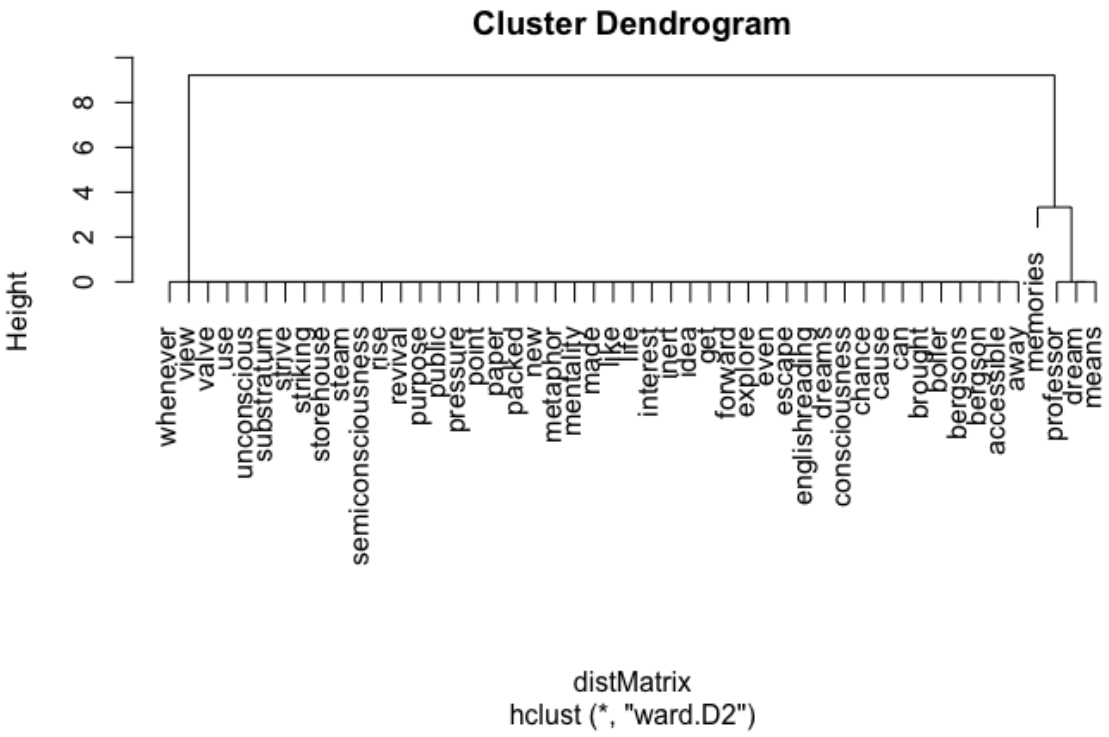
Paragraph 1:



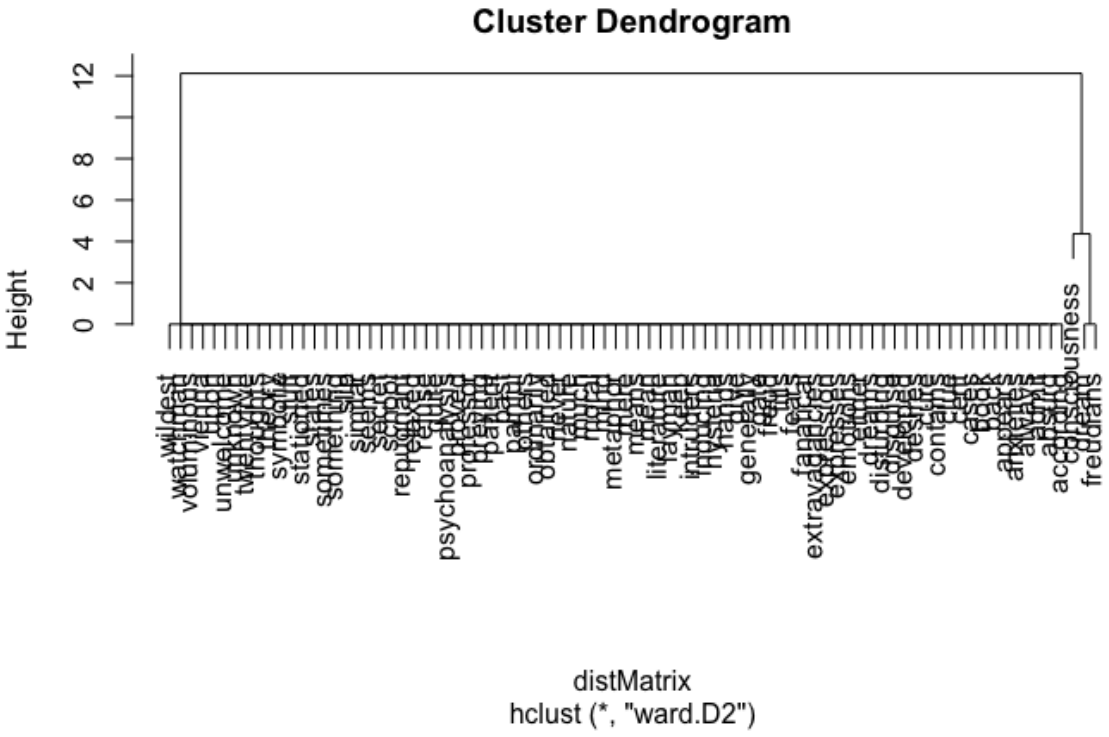
Paragraph 2:



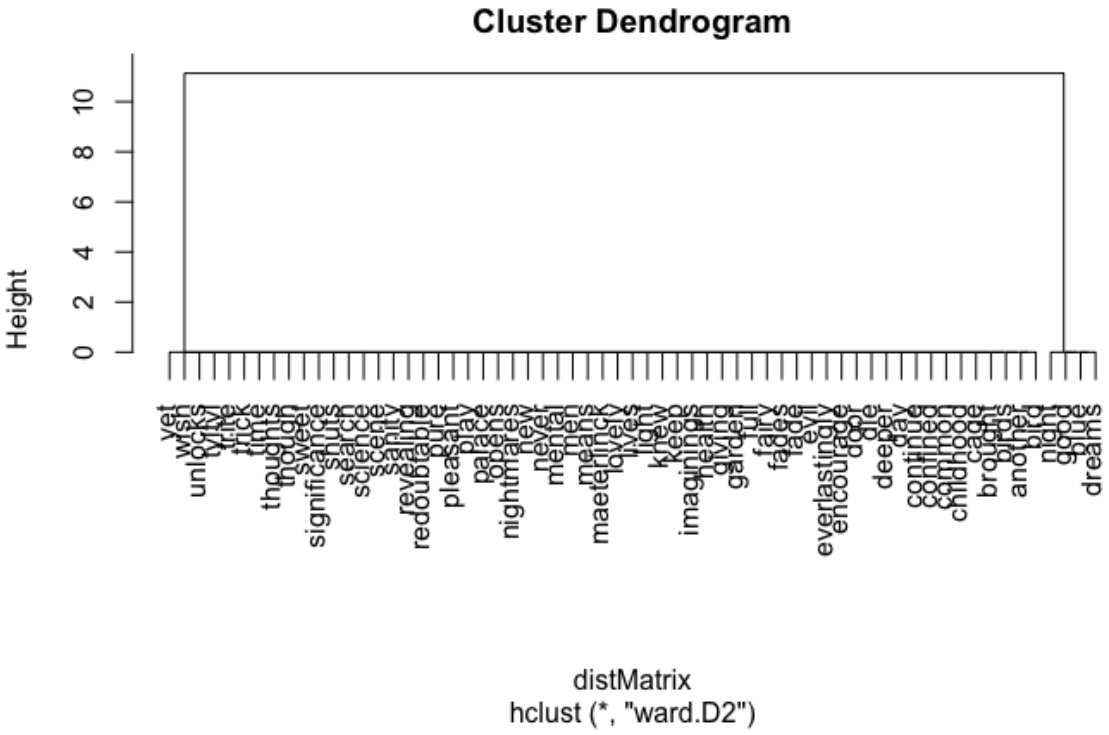
Paragraph 3:



Paragraph 4:

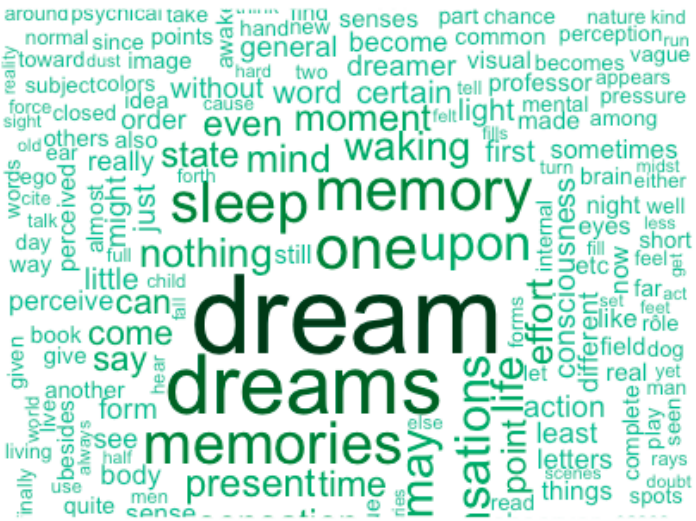


Paragraph 7:



(2) WordCloud:

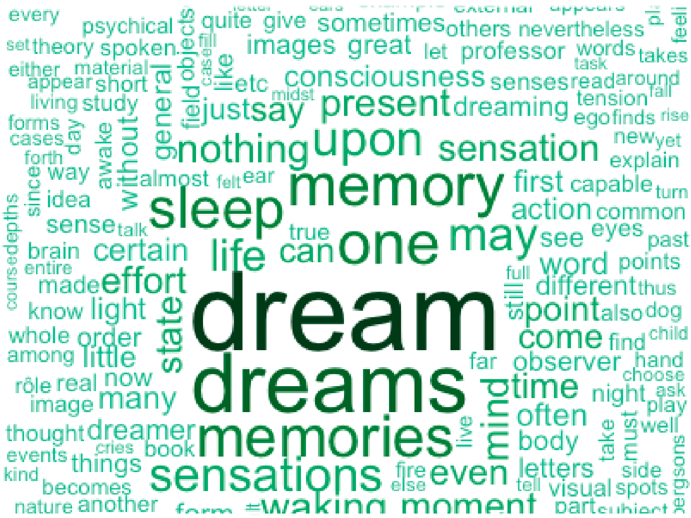
Paragraph 1:



Paragraph 2:

Paragraph 3:

Paragraph 4:


dream
dreams
memory
memories
sleep
one
may
upon
nothing
sensation
sensations
present
consciousness
effort
state
life
can
say
just
dreaming
images
great
sometimes
first
action
word
different
point
come
time
often
body
even
letters
mind
many
dreamer
light
little
order
certain

Paragraph 7:


dream
dreams
memories
memory
sleep
one
may
sensations
moment
upon
present
life
nothing
state
waking
mind
point
come
consciousness
dreaming
images
time
effort
even
can
say
light
certain
form
general
many

2.d Longest word and sentences

In order to find the longest word, the longest sentence, the length of the longest sentence and the shortest sentence in each paragraph, firstly we need to separate the text into paragraphs.

```{r}
# Chunk txt into paragraphs
chunk_into_paragraphs <- function(text) {
  break_points <- c(1, as.numeric(gregexpr("\"\"", text)[[1]]) + 1)
  paragraphs <- NULL
  for(i in 1:length(break_points)) {
    res <- substr(text, break_points[i], break_points[i+1])
    if(i>1) { paragraphs[i] <- sub('. ', '', res) } else { paragraphs[i] <- res }
    }
  paragraphs <- paragraphs[paragraphs=!is.na(paragraphs)]
  removeInd <- NULL
  for(i in 1:length(paragraphs)){
    if(count_words(paragraphs[i])<7){
      removeInd <-append(removeInd,i)
    }
    }
  for(i in 1:length(removeInd)){
    paragraphs <- paragraphs[-(removeInd[i]-i+1)]
    }
  return(paragraphs)
}
mycorpus <- VCorpus(VectorSource(SATtxt))

corpus_frame <- data.frame(text=unlist(sapply(mycorpus, `[`, "content")), stringsAsFactors=F)
paragraphs <-chunk_into_paragraphs(corpus_frame)

paragraphs
```

This function is used to generate a vector which contains the paragraphs of the input text. To locate each of the paragraphs, we use “\”\”” to separate texts into paragraphs. If the number of words no less than 7, we save it into a vector.

From 2.b, we can easily find the longest sentence. Similarly, we use the same method to find the shortest sentence.

```
# Project_2 1.(d)
Find_shortest_sentence <- function(text){
  result<-NULL
  wordCount <-NULL
  for(i in 1:length(text)){
    wordCount[i] = count_words(text[i])
  }
  min <-min(wordCount)
  for(i in 1:length(wordCount)){
    if(wordCount[i] == min)
      index <-i
  }
  return(index)
}

sentence <-chunk_into_sentences(paragraphs[1])
#sentence

Find_longest_sentence(sentence)
sentence[Find_longest_sentence(sentence)]
count_words(sentence[Find_longest_sentence(sentence)])

Find_shortest_sentence(sentence)
sentence[Find_shortest_sentence(sentence)]
count_words(sentence[Find_shortest_sentence(sentence)])
```

To find the longest word, we count the character number for each word and find the word with the most characters.

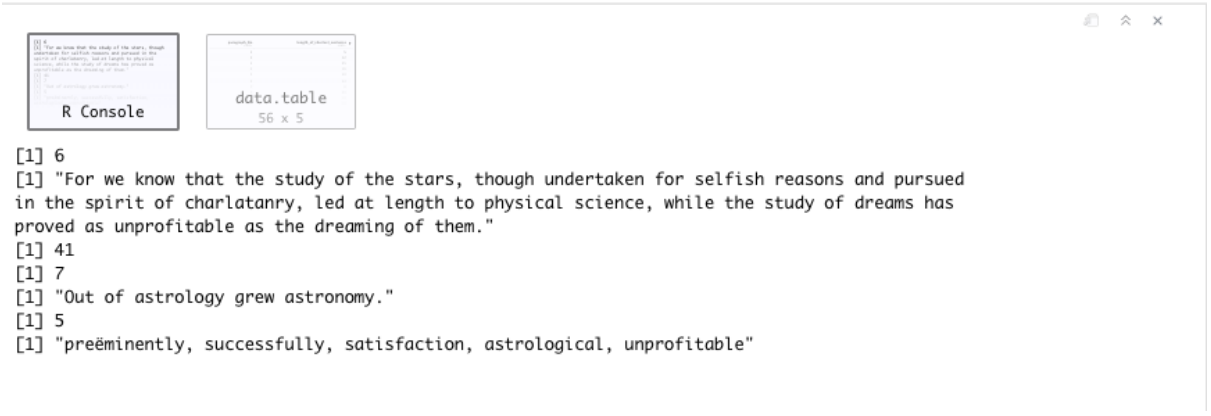
```
Find_longest_word <- function(text){
  word <-tokenize_words(text)
  charCount <- NULL
  result <- NULL

  for(i in 1:length(word[[1]])){
    charCount[i] = count_characters(word[[1]][i])
  }
  max <-max(charCount)
  for(i in 1:length(charCount)){
    if(charCount[i] == max){
      result <-append(result,word[[1]][i])
    }
  }
  return(toString(result))
}
Find_longest_word(paragraphs[1])
```

After that, we use a datatable to display the required variable in all paragraphs.

```
dt<-data.table(paragraph_No = 1, length_of_shortest_sentence = count_words(sentence[Find_shortest_sentence(sentence)]),
               length_of_longest_sentence = count_words(sentence[Find_longest_sentence(sentence)]),
               longest_word = Find_longest_word(paragraphs[1]),longest_sentence =
sentence[Find_longest_sentence(sentence)])
for(i in 2:length(paragraphs)){
  sentence <-chunk_into_sentences(paragraphs[i])

  dt1<-data.table(paragraph_No = i, length_of_shortest_sentence =
count_words(sentence[Find_shortest_sentence(sentence)]),
               length_of_longest_sentence = count_words(sentence[Find_longest_sentence(sentence)]),
               longest_word = Find_longest_word(paragraphs[i]),longest_sentence =
sentence[Find_longest_sentence(sentence)])
  dt<-merge(dt, dt1, all = TRUE)
}
print(dt)
```

The result is shown as below:

(1) As an example, I display the deliverariables in the first paragraph.



It shows that in the first paragraph, the longest sentence is the 6th sentence and it contains 41 words.

"For we know that the study of the stars, though undertaken for selfish reasons and pursued in the spirit of charlatanry, led at length to physical science, while the study of dreams has proved as unprofitable as the dreaming of them."

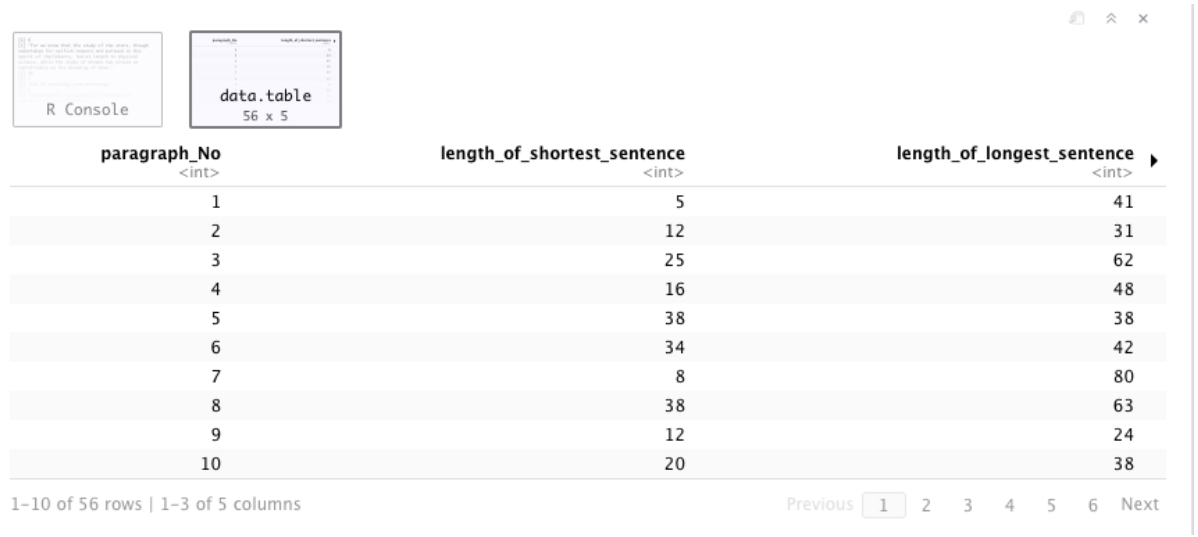
The shortest sentence is the 7th sentence and it contains 5 words.

"Out of astrology grew astronomy."

There are totally fix words containing the most characters(12 characters).

"preëminently, successfully, satisfaction, astrological, unprofitable"

(2) Here is the table which display the required deliverariables. Totally there are 56 paragraphs in the text.

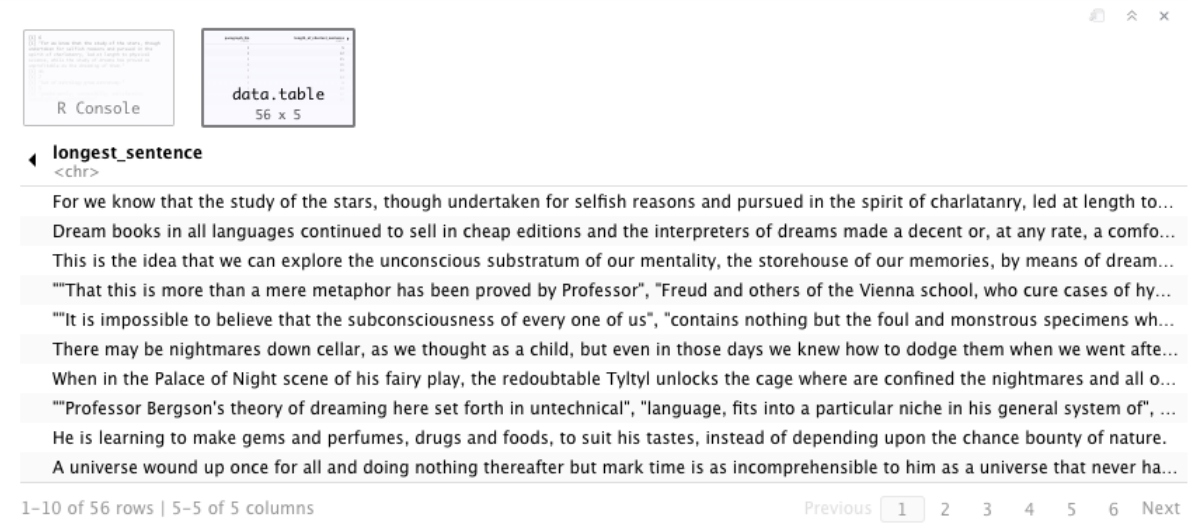


paragraph_No <int>	length_of_shortest_sentence <int>	length_of_longest_sentence <int>
1	5	41
2	12	31
3	25	62
4	16	48
5	38	38
6	34	42
7	8	80
8	38	63
9	12	24
10	20	38

1–10 of 56 rows | 1–3 of 5 columns Previous 1 2 3 4 5 6 Next

longest_word <chr>
preëminently, successfully, satisfaction, astrological, unprofitable
substantially
consciousness, consciousness
extravagancies
subconsciousness
unpleasant, nightmares
everlastingly
contemporary
descriptive, consciously
incomprehensible

1–10 of 56 rows | 4–4 of 5 columns Previous 1 2 3 4 5 6 Next



2.e WordNet Demonstration

"WordNet" package is used to get a synonym of a word from the dict. Here is an example. The antonym of "hot" is cold.

```
setDict("/Users/tianyuyang/Desktop/WordNet-3.0/dict")
Sys.setenv(WNHOME = "/Users/tianyuyang/Desktop/WordNet-3.0/")
syn_list  <- apply(wnet,by=1,function(row){synonyms(row["word"],row["ss_type"])})

if(initDict()) {
  filter <- getTermFilter("ExactMatchFilter", "hot", TRUE)
  terms <- getIndexTerms("ADJECTIVE", 5, filter)
  synsets <- getSynsets(terms[[1]])
  related <- getRelatedSynsets(synsets[[1]], "!")
  sapply(related, getWord)
}
```

```
[1] "cold"
```

This is to get the nouns related to the word "car".

```
  filter <- getTermFilter("StartsWithFilter", "car", TRUE)
  terms <- getIndexTerms("NOUN", 5, filter)
  sapply(terms, getLemma)
}
```

```
[1] "car"          "car-ferry"    "car-mechanic" "car battery"  "car bomb"
```

To get verbs and nouns in the first five paragraphs which contain no less than 5 characters, we firstly need to get the first five paragraphs. We use the paragraph function in 2.d.

```
# Create text for the first five paragraphs
firstFive_Paragraph <-NULL
for(i in 1:5){
  firstFive_Paragraph <-append(firstFive_Paragraph,paragraphs[i])
}
firstFive_Paragraph <-toString(firstFive_Paragraph)
firstFive_Paragraph <-removeNumPunct(firstFive_Paragraph)
firstFive_Paragraph
```

```
[1] "Before the dawn of history mankind was engaged in the study of dreaming The wise man among the ancients was preëminently the interpreter of
dreams The ability to interpret successfully or plausibly was the quickest road to royal favor as Joseph and Daniel found it to be failure to give
satisfaction in this respect led to banishment from court or death When a scholar laboriously translates a cuneiform tablet dug up from a
Babylonian mound where it has lain buried for five \nthousand years or more the chances are that it will turn out either an astrological treatise
or a dream book If the former we look upon it with some indulgence if the latter with pure contempt For we know that the study of the stars though
undertaken for selfish reasons and pursued in the spirit of charlatanry led at length to physical science while the study of dreams has proved as
unprofitable as the dreaming of them Out of astrology grew astronomy Out of oneiromancy has \ngrownnothing  That at least was substantially true
up to the beginning of the present century Dream books in all languages continued to sell in cheap editions and the interpreters of dreams made a
decent or at any rate a comfortable living out of the poorer classes But the psychologist rarely paid attention to dreams except incidentally in
his study of imagery association and the speed of thought But now a change has come over the spirit of the times The subject of the significance
of dreams \nso long ignored has suddenly become a matter of energetic study and of fiery controversy the world over  The cause of this revival of
interest is the new point of view brought forward by Professor Bergson in the paper which is here made accessible to the Englishreading public
This is the idea that we can explore the unconscious substratum of our mentality the storehouse of our memories by means of dreams for these
memories are by no means inert but have as it were a life and purpose of their own and strive to rise into \nconsciousness whenever they get a
chance even into the semiconsciousness of a dream To use Professor Bergsons striking metaphor our memories are packed away under pressure like
steam in a boiler and the dream is their escape valve  That this is more than a mere metaphor has been proved by Professor Freud and others of the
Vienna school who cure cases of hysteria by inducing the patient to give expression to the secret anxieties and emotions which unknown to him have
been preying upon his mind The \nclue to these disturbing thoughts is generally obtained in dreams or similar states of relaxed consciousness
According to the Freudians a dream always means something but never what it appears to mean It is symbolic and expresses desires or fears which we
refuse ordinarily to admit to consciousness either because they are painful or because they are repugnant to our moral nature A watchman is
stationed at the gate of consciousness to keep them back but sometimes these unwelcome \nintruders slip past him in disguise In the hands of
fanatical Freudians this theory has developed the wildest extravagancies and the voluminous literature of psychoanalysis contains much that seems
to the layman quite as absurd as the stuff which fills the twentyfive cent dream book  It is impossible to believe that the subconsciousness of
every one of us contains nothing but the foul and monstrous specimens which they dredge up from the mental depths of their neuropathic patients
and exhibit with \nsuch pride "
```

Get the words that contain no less than 5 characters and make them as a list.

```
# Get words which its length no less than 5
Five_words <-tokenize_words(firstFive_Paragraph)[[1]]
#Five_words
word_list<-NULL
for(i in 1:length(Five_words)){
  if(length(tokenize_characters(Five_words[i])[[1]])>=5){
    word_list <-append(word_list,Five_words[i])
  }
}
word_list
```

```
  [1] "before"      "history"     "mankind"     "engaged"     "study"       "dreaming"
  [7] "among"       "ancients"    "preëminently" "interpreter" "dreams"      "ability"
 [13] "interpret"   "successfully" "plausibly"  "quickest"    "royal"       "favor"
 [19] "joseph"      "daniel"      "found"       "failure"     "satisfaction" "respect"
 [25] "banishment"  "court"       "death"       "scholar"     "laboriously" "translates"
 [31] "cuneiform"   "tablet"      "babylonian"  "mound"       "where"       "buried"
 [37] "thousand"    "years"       "chances"     "either"      "astrological" "treatise"
```

Use synonyms(word, pos) function in wordnet to judge the part of speech of a word. pos is the input part of speech, including “NOUN”, “VERB”, “ADVERB” and “ADJECTIVE”. To get all the verbs and nouns. We set pos as NOUN/VERB and judge the length of synoyms(word, pos). If the length is 0, it is not a noun/verb. If the length is non-zero, it is a noun/verb.

The noun list for the first five paragraphs that contain no less than five characters is:

```
verb_list<-NULL
noun_list<-NULL
for(i in 1:length(word_list)){
  if(length(synonyms(word_list[i],"NOUN"))!=0){
    noun_list <-append(noun_list,word_list[i])
  }
}
for(i in 1:length(word_list)){
  if(length(synonyms(word_list[i],"VERB"))!=0){
    verb_list <-append(verb_list,word_list[i])
  }
}
noun_list
verb_list
```

```
  [1] "history"          "mankind"          "study"            "dreaming"         "ancients"         "interpreter"
  [7] "ability"          "royal"            "favor"            "joseph"           "daniel"           "found"
 [13] "failure"          "satisfaction"     "respect"          "banishment"       "court"            "death"
 [19] "scholar"          "cuneiform"        "tablet"           "babylonian"       "mound"            "thousand"
 [25] "years"            "treatise"         "dream"            "former"           "indulgence"       "latter"
 [31] "contempt"         "study"            "pursued"          "spirit"           "length"           "science"
 [37] "while"            "study"            "dreaming"         "astrology"        "astronomy"        "oneiromancy"
 [43] "least"            "beginning"        "present"          "century"          "dream"            "living"
 [49] "psychologist"     "attention"        "study"            "imagery"          "association"      "speed"
 [55] "thought"          "change"           "spirit"           "times"            "subject"          "significance"
 [61] "matter"           "study"            "controversy"      "world"            "cause"            "revival"
 [67] "interest"         "point"            "forward"          "professor"        "bergson"          "paper"
 [73] "public"           "unconscious"      "substratum"       "mentality"        "storehouse"       "means"
 [79] "means"            "purpose"          "consciousness"    "chance"           "semiconsciousness" "dream"
 [85] "professor"        "striking"         "metaphor"         "pressure"         "steam"            "boiler"
 [91] "dream"            "escape"           "valve"            "metaphor"         "professor"        "freud"
 [97] "vienna"           "school"           "hysteria"         "inducing"         "patient"          "expression"
[103] "secret"           "unknown"          "consciousness"    "dream"            "means"            "refuse"
[109] "consciousness"    "moral"            "nature"           "watchman"         "consciousness"    "disguise"
[115] "hands"            "theory"           "literature"       "psychoanalysis"   "layman"           "absurd"
[121] "stuff"            "dream"            "impossible"       "subconsciousness" "nothing"          "dredge"
[127] "exhibit"          "pride"
```

The noun list for the first five paragraphs that contain no less than five characters is:

```
verb_list<-NULL
noun_list<-NULL
for(i in 1:length(word_list)){
  if(length(synonyms(word_list[i],"NOUN"))!=0){
    noun_list <-append(noun_list,word_list[i])
  }
}
for(i in 1:length(word_list)){
  if(length(synonyms(word_list[i],"VERB"))!=0){
    verb_list <-append(verb_list,word_list[i])
  }
}
noun_list
verb_list
```

```
 [1] "study"     "interpret" "favor"     "found"     "respect"   "court"     "mound"     "dream"     "study"     "spirit"    "study"
[12] "present"   "dream"     "except"    "study"     "speed"     "change"    "spirit"    "subject"   "become"    "matter"    "study"
[23] "cause"     "interest"  "point"     "forward"   "paper"     "explore"   "purpose"   "strive"    "chance"    "dream"     "pressure"
[34] "steam"     "dream"     "escape"    "school"    "dream"     "refuse"    "admit"     "disguise"  "stuff"     "dream"     "believe"
[45] "dredge"    "exhibit"   "pride"     "study"     "interpret" "favor"     "found"     "respect"   "court"     "mound"     "dream"
[56] "study"     "spirit"    "study"     "present"   "dream"     "except"    "study"     "speed"     "change"    "spirit"    "subject"
[67] "become"    "matter"    "study"     "cause"     "interest"  "point"     "forward"   "paper"     "explore"   "purpose"   "strive"
[78] "chance"    "dream"     "pressure"  "steam"     "dream"     "escape"    "school"    "dream"     "refuse"    "admit"     "disguise"
[89] "stuff"     "dream"     "believe"   "dredge"    "exhibit"   "pride"
```

2.f Word Frequency Analysis

To use zipfR package to analyze word frequency, we firstly need to obtain the right data structure for the input word frequency.

```
# Project_2 1.(f)
# Corpus Management
setwd("/Users/tianyuyang/Desktop/project2/CSCE6444_Project2")
SAT<-VCorpus(DirSource(".", ignore.case=TRUE, mode="text"))
SATlow <- tm_map(SAT[2], content_transformer(tolower))
# Remove numbers, punctuation and stopwords
SATcl <- tm_map(SATlow,content_transformer(removeNumPunct))
myStopwords <- c(stopwords('english'))
SATstop <- tm_map(SATcl, removeWords, myStopwords)

SATtdm<-TermDocumentMatrix(SATstop, control = list(wordLengths = c(1,Inf)))

# Term Frequency
term.freq <-rowSums(as.matrix(SATtdm))
# Frequency no less than 1
term.freq <-subset(term.freq,term.freq>=1)
df <-data.frame(term = names(term.freq), f = term.freq)
setwd("/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/word_frequency/")
write.table(df, "word_frequency.txt", append = FALSE, sep = "\t",
            row.names = FALSE, col.names = TRUE)
setwd("/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/word_frequency/")
SAT<-VCorpus(DirSource(".", ignore.case=TRUE, mode="text"))
rem <-removePunctuation(SAT[[1]]$content)
writeLines(rem,"word_frequency.txt")
txttfl <-read.tfl("word_frequency.txt")
txttfl
```

```
summary(txttfl)
```

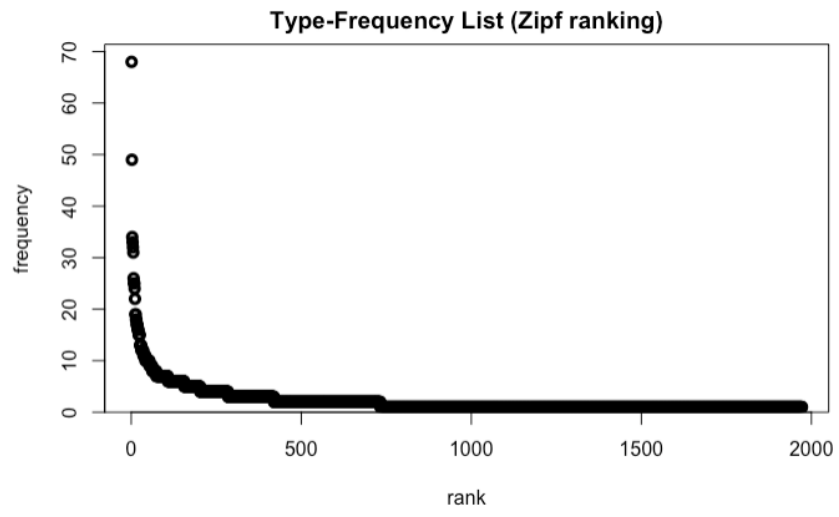
```
zipfR object for frequency spectrum
Sample size:     N  = 4440
Vocabulary size: V  = 1972
Range of freq's: f  = 1 ... 68
Mean / median:   mu = 2.251521 ,  M = 1
Hapaxes etc.:    V1 = 1240 ,  V2 = 313
```

```
summary(txt.spc)
```

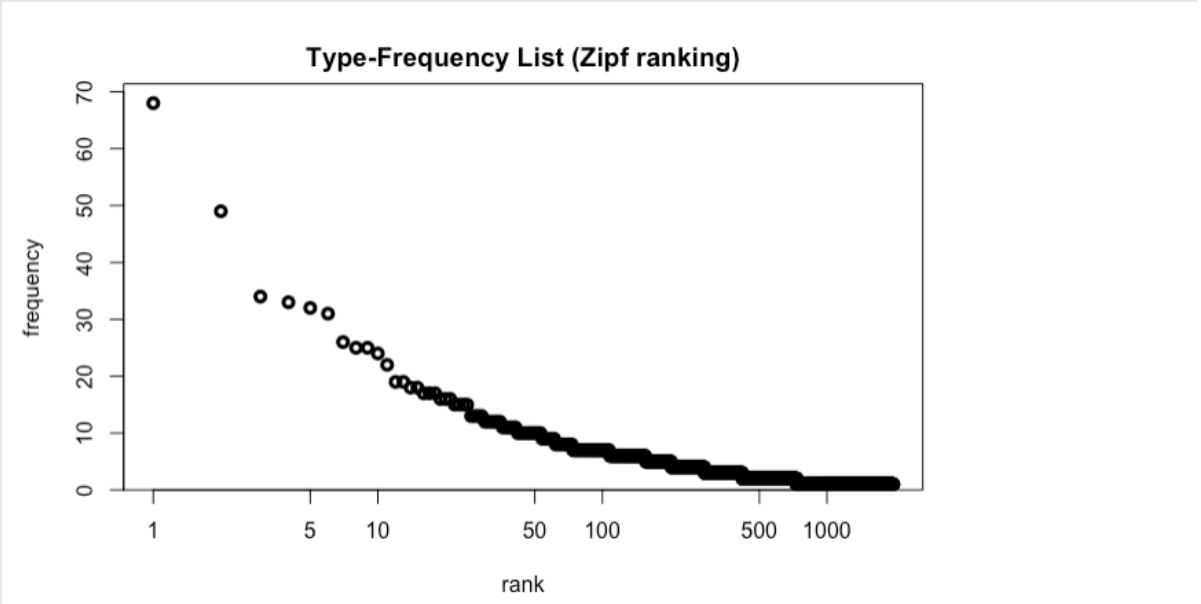
```
zipfR object for frequency spectrum
Sample size:     N  = 4440
Vocabulary size: V  = 1972
Class sizes:     Vm = 1240 313 135 82 47 48 34 12 ...
```

(1) Plot the spectrum

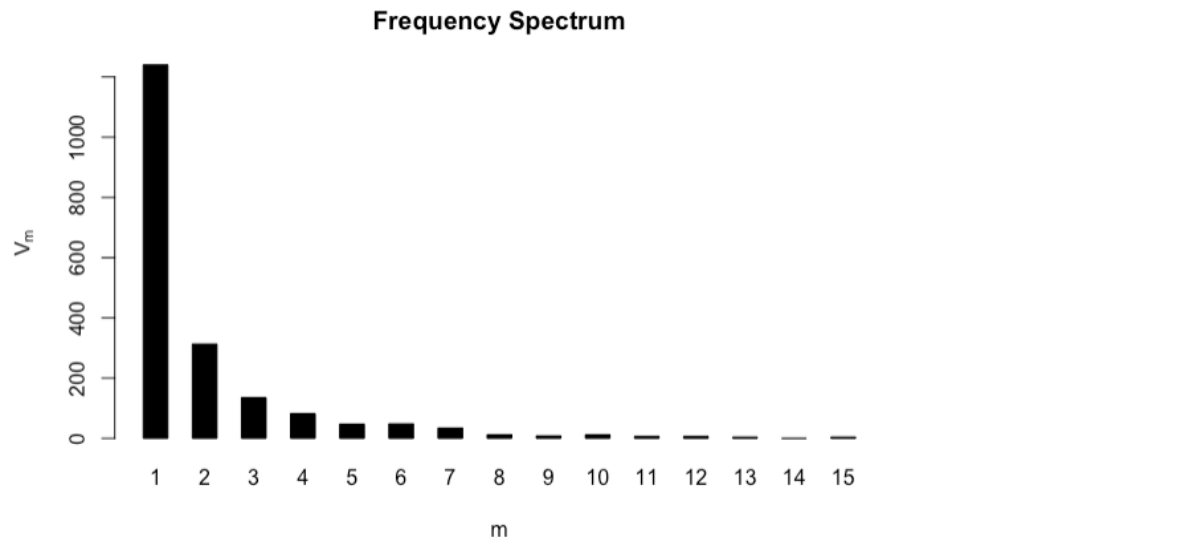
```
plot(txttfl)
```



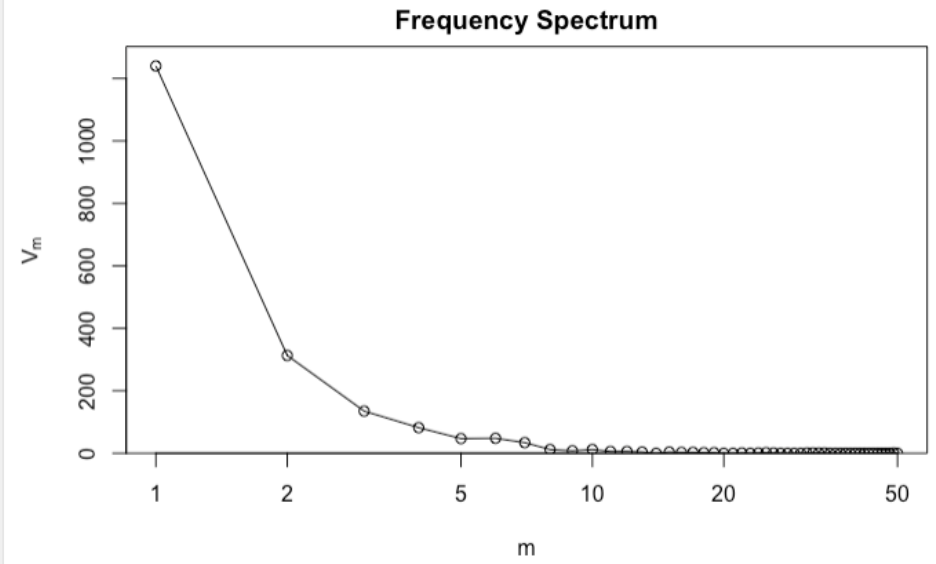
```
plot(txttfl, log="x")
```



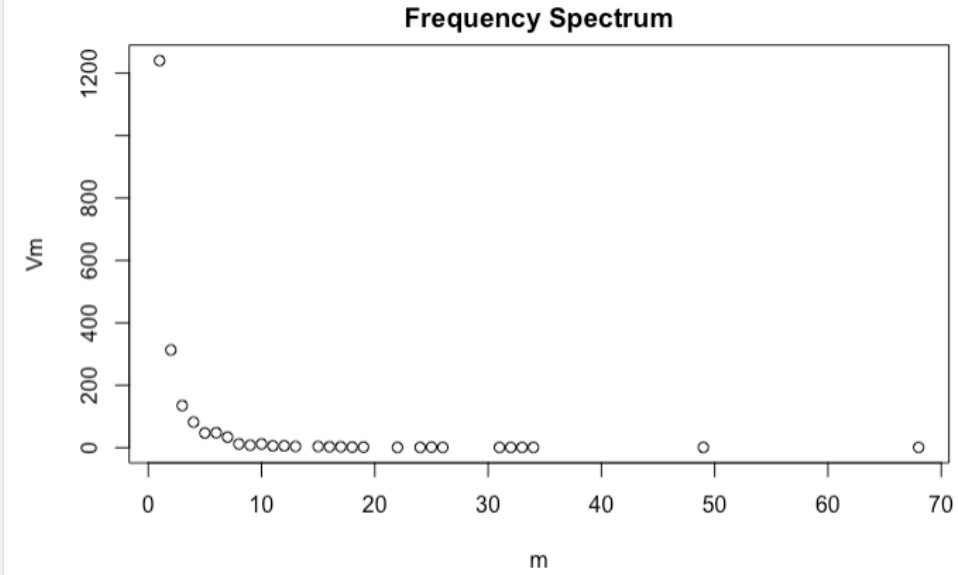

```
plot(txt.spc)
```



```
plot(txt.spc,log="x")
```




```
with(txt.spc, plot(m, Vm, main="Frequency Spectrum"))
```

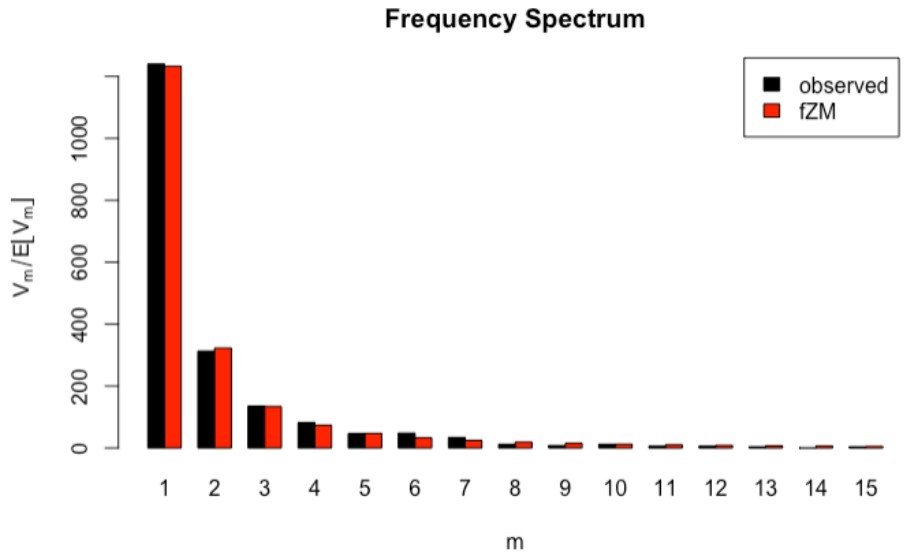


(2) Estimating V and other quantities at arbitrary sample sizes

```
txt.fzm <- lnre("fzm",txt.spc)
summary(txt.fzm)
```


```
finite Zipf-Mandelbrot LNRE model.
Parameters:
   Shape:          alpha = 0.7999539
   Lower cutoff:       A = 1.566091e-05
   Upper cutoff:       B = 0.004305149
 [ Normalization:      C = 0.8814947 ]
Population size: S = 7606.641
Sampling method: Poisson, with exact calculations.

Parameters estimated from sample of size N = 4440:

Goodness-of-fit (multivariate chi-squared test):
```

```
plot(txt.spc, txt.fzm.spc, legend=c("observed", "fZM"))
```



(3) Vocabulary growth curves

```
head(txt.vgc)
```

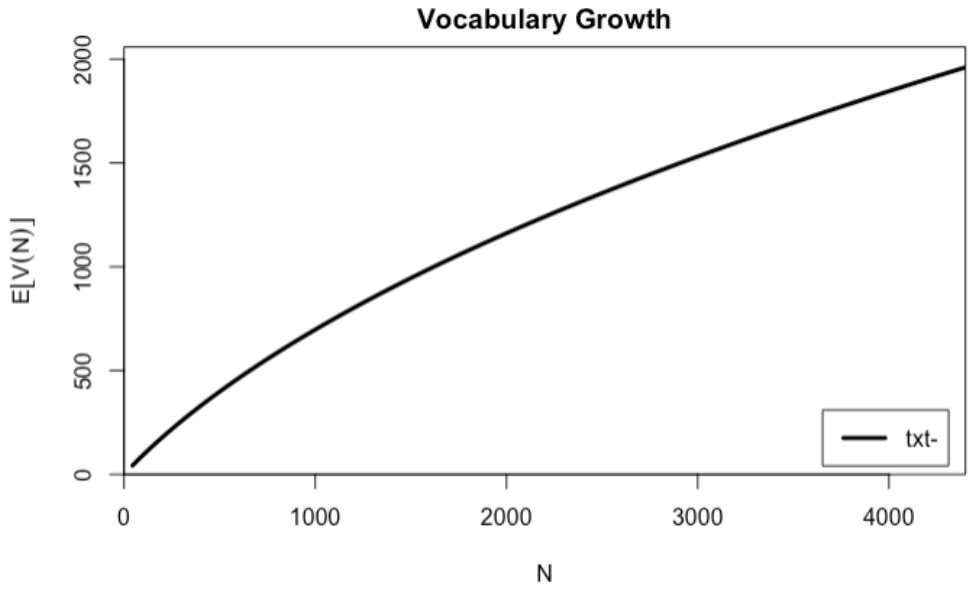
```
    N         V
1  44  42.72071
2  88  83.22594
3 132 121.90543
4 176 159.03422
5 220 194.81468
6 264 229.40152
```

```
txt.vgc
```

```
       N         V
3    132  121.9054
5    220  194.8147
18   792  580.9807
20   880  631.4513
24  1056  727.4106
25  1100  750.4656
26  1144  773.1772
27  1188  795.5590
30  1320  860.8491
34  1496  943.9775
36  1584  984.0286
37  1628 1003.7046
41  1804 1080.2437
42  1848 1098.8695
43  1892 1117.3037
47  2068 1189.2288
49  2156 1224.1722
58  2552 1374.0221
65  2860 1483.3094
71  3124 1572.6753
74  3256 1616.0166
87  3828 1794.7485
89  3916 1821.0591
93  4092 1872.8198
100 4400 1960.8033
```

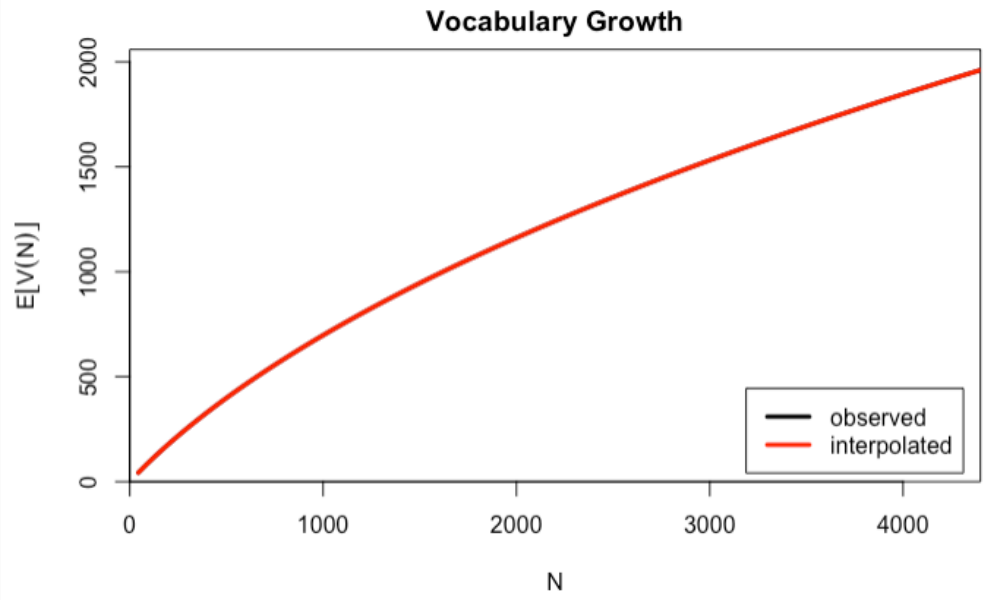
```
summary(txt.vgc)
```

```
zipfR object for expected vocabulary growth curve
100 samples for N = 44 ... 4400
```



(4) Interpolation

```
# Interpolation
txt.bin.vgc <- vgc.interp(txt.spc, N(txt.vgc), m.max=1)
plot(txt.vgc, txt.bin.vgc,legend=c("observed", "interpolated"))
```



2.g Bigrams and Trigrams

To obtain bigrams and trigrams in the first three paragraphs, we firstly need to get a string which contains the content of the first three paragraphs and then remove all the numbers and punctuation.

```{r}
# Project_2 1.(g)
# Create text for the first three paragraphs
for(i in 1:3){
  firstThree_Paragraph <-append(firstThree_Paragraph,paragraphs[i])
}
firstThree_Paragraph <-toString(firstThree_Paragraph)
firstThree_Paragraph <-removeNumPunct(firstThree_Paragraph)

SAT1tokens<-tokens(firstThree_Paragraph)

SATbigrams <- tokens_ngrams(SAT1tokens, n = 2L, skip = 0L, concatenator = "_")
SATbigrams
SATtrigrams <- tokens_ngrams(SAT1tokens, n = 3L, skip = 0L, concatenator = "_")
SATtrigrams
```

The bigrams are shown as below:

```
tokens from 1 document.
text1 :
   [1] "Before_the"            "the_dawn"              "dawn_of"               "of_history"            "history_mankind"
   [6] "mankind_was"           "was_engaged"           "engaged_in"            "in_the"                "the_study"
  [11] "study_of"              "of_dreaming"           "dreaming_The"          "The_wise"              "wise_man"
  [16] "man_among"             "among_the"             "the_ancients"          "ancients_was"          "was_preëminently"
  [21] "preëminently_the"      "the_interpreter"       "interpreter_of"        "of_dreams"             "dreams_The"
  [26] "The_ability"           "ability_to"            "to_interpret"          "interpret_successfully" "successfully_or"
  [31] "or_plausibly"          "plausibly_was"         "was_the"               "the_quickest"          "quickest_road"
  [36] "road_to"               "to_royal"              "royal_favor"           "favor_as"              "as_Joseph"
  [41] "Joseph_and"            "and_Daniel"            "Daniel_found"          "found_it"              "it_to"
  [46] "to_be"                 "be_failure"            "failure_to"            "to_give"               "give_satisfaction"
  [51] "satisfaction_in"       "in_this"               "this_respect"          "respect_led"           "led_to"
  [56] "to_banishment"         "banishment_from"       "from_court"            "court_or"              "or_death"
  [61] "death_When"            "When_a"                "a_scholar"             "scholar_laboriously"   "laboriously_translates"
  [66] "translates_a"          "a_cuneiform"           "cuneiform_tablet"      "tablet_dug"            "dug_up"
  [71] "up_from"               "from_a"                "a_Babylonian"          "Babylonian_mound"      "mound_where"
  [76] "where_it"              "it_has"                "has_lain"              "lain_buried"           "buried_for"
  [81] "for_five"              "five_thousand"         "thousand_years"        "years_or"              "or_more"
  [86] "more_the"              "the_chances"           "chances_are"           "are_that"              "that_it"
  [91] "it_will"               "will_turn"             "turn_out"              "out_either"            "either_an"
  [96] "an_astrological"       "astrological_treatise" "treatise_or"           "or_a"                  "a_dream"
 [101] "dream_book"            "book_If"               "If_the"                "the_former"            "former_we"
 [106] "we_look"               "look_upon"             "upon_it"               "it_with"               "with_some"
 [111] "some_indulgence"       "indulgence_if"         "if_the"                "the_latter"            "latter_with"
 [116] "with_pure"             "pure_contempt"         "contempt_For"          "For_we"                "we_know"
 [121] "know_that"             "that_the"              "the_study"             "study_of"              "of_the"
 [126] "the_stars"             "stars_though"          "though_undertaken"     "undertaken_for"        "for_selfish"
 [131] "selfish_reasons"       "reasons_and"           "and_pursued"           "pursued_in"            "in_the"
 [136] "the_spirit"            "spirit_of"             "of_charlatanry"        "charlatanry_led"       "led_at"
 [141] "at_length"             "length_to"             "to_physical"           "physical_science"      "science_while"
 [146] "while_the"             "the_study"             "study_of"              "of_dreams"             "dreams_has"
 [151] "has_proved"            "proved_as"             "as_unprofitable"       "unprofitable_as"       "as_the"
 [156] "the_dreaming"          "dreaming_of"           "of_them"               "them_Out"              "Out_of"
 [161] "of_astrology"          "astrology_grew"        "grew_astronomy"        "astronomy_Out"         "Out_of"
 [166] "of_oneiromancy"        "oneiromancy_has"       "has_grownnothing"      "grownnothing_That"     "That_at"
 [171] "at_least"              "least_was"             "was_substantially"     "substantially_true"    "true_up"
 [176] "up_to"                 "to_the"                "the_beginning"         "beginning_of"          "of_the"
```

The trigrams are shown as below:

```
tokens from 1 document.
text1 :
   [1] "Before_the_dawn"                "the_dawn_of"                "dawn_of_history"                "of_history_mankind"
   [5] "history_mankind_was"            "mankind_was_engaged"        "was_engaged_in"                 "engaged_in_the"
   [9] "in_the_study"                   "the_study_of"               "study_of_dreaming"              "of_dreaming_The"
  [13] "dreaming_The_wise"              "The_wise_man"               "wise_man_among"                 "man_among_the"
  [17] "among_the_ancients"             "the_ancients_was"           "ancients_was_preëminently"      "was_preëminently_the"
  [21] "preëminently_the_interpreter"   "the_interpreter_of"         "interpreter_of_dreams"          "of_dreams_The"
  [25] "dreams_The_ability"             "The_ability_to"             "ability_to_interpret"           "to_interpret_successfully"
  [29] "interpret_successfully_or"      "successfully_or_plausibly"  "or_plausibly_was"              "plausibly_was_the"
  [33] "was_the_quickest"               "the_quickest_road"          "quickest_road_to"               "road_to_royal"
  [37] "to_royal_favor"                 "royal_favor_as"             "favor_as_Joseph"                "as_Joseph_and"
  [41] "Joseph_and_Daniel"              "and_Daniel_found"           "Daniel_found_it"                "found_it_to"
  [45] "it_to_be"                       "to_be_failure"              "be_failure_to"                  "failure_to_give"
  [49] "to_give_satisfaction"           "give_satisfaction_in"       "satisfaction_in_this"           "in_this_respect"
  [53] "this_respect_led"               "respect_led_to"             "led_to_banishment"              "to_banishment_from"
  [57] "banishment_from_court"          "from_court_or"              "court_or_death"                 "or_death_When"
  [61] "death_When_a"                   "When_a_scholar"             "a_scholar_laboriously"          "scholar_laboriously_translates"
  [65] "laboriously_translates_a"       "translates_a_cuneiform"     "a_cuneiform_tablet"             "cuneiform_tablet_dug"
  [69] "tablet_dug_up"                  "dug_up_from"                "up_from_a"                      "from_a_Babylonian"
  [73] "a_Babylonian_mound"             "Babylonian_mound_where"     "mound_where_it"                 "where_it_has"
  [77] "it_has_lain"                    "has_lain_buried"            "lain_buried_for"                "buried_for_five"
  [81] "for_five_thousand"              "five_thousand_years"        "thousand_years_or"              "years_or_more"
  [85] "or_more_the"                    "more_the_chances"           "the_chances_are"                "chances_are_that"
  [89] "are_that_it"                    "that_it_will"               "it_will_turn"                   "will_turn_out"
  [93] "turn_out_either"                "out_either_an"              "either_an_astrological"         "an_astrological_treatise"
  [97] "astrological_treatise_or"       "treatise_or_a"              "or_a_dream"                     "a_dream_book"
 [101] "dream_book_If"                  "book_If_the"                "If_the_former"                  "the_former_we"
 [105] "former_we_look"                 "we_look_upon"               "look_upon_it"                   "upon_it_with"
 [109] "it_with_some"                   "with_some_indulgence"       "some_indulgence_if"             "indulgence_if_the"
 [113] "if_the_latter"                  "the_latter_with"            "latter_with_pure"               "with_pure_contempt"
 [117] "pure_contempt_For"              "contempt_For_we"            "For_we_know"                    "we_know_that"
 [121] "know_that_the"                  "that_the_study"             "the_study_of"                   "study_of_the"
 [125] "of_the_stars"                   "the_stars_though"           "stars_though_undertaken"        "though_undertaken_for"
 [129] "undertaken_for_selfish"         "for_selfish_reasons"        "selfish_reasons_and"            "reasons_and_pursued"
 [133] "and_pursued_in"                 "pursued_in_the"             "in_the_spirit"                  "the_spirit_of"
 [137] "spirit_of_charlatanry"          "of_charlatanry_led"         "charlatanry_led_at"             "led_at_length"
 [141] "at_length_to"                   "length_to_physical"         "to_physical_science"            "physical_science_while"
```

2.h Three Packages Demonstration

2.h.1 Three packages

(1)CorpusTools:

1.laplace

To laplace (i.e. add constant) smoothing a numeric vector of term frequencies.

```
> #laplace in corpustools
> lapfreq <- laplace(freq, add = 0.5)
> lapfreq
     ability          able      abnormal     abolition        abound     abruptly        absent       absorbs       abstain
0.0002792776 0.0006516477 0.0002792776 0.0006516477 0.0002792776 0.0004654627 0.0002792776 0.0002792776 0.0002792776
    abstract        absurd     absurdity     abundance   acceleration    acceptance      accepted       accepts    accessible
0.0002792776 0.0006516477 0.0004654627 0.0004654627 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0002792776
   accidents    accompanies    accomplish   accomplished     accordance      according       account      accounts   accumulated
0.0002792776 0.0002792776 0.0006516477 0.0004654627 0.0002792776 0.0006516477 0.0002792776 0.0002792776 0.0002792776
  acquainted      acquiring            act         acting         action      actionin         active       activity         actual
0.0002792776 0.0002792776 0.0006516477 0.0004654627 0.0023273133 0.0002792776 0.0002792776 0.0002792776 0.0002792776
    actually         acuity          adapt     adaptation        adapted            add        address          adds     adherence
0.0004654627 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0002792776 0.0002792776
       adjust     adjustment        adjusts     admission          admit        adopted      advantage    advantageous          afar
0.0002792776 0.0008378328 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0002792776 0.0004654627
       affect       affected      affection     affections      affective           ages     aggravates          ago        agree
0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0002792776
    agreeable          aid          air          airy         ajar         akin         ālan    alexandria       alfred
0.0002792776 0.0002792776 0.0004654627 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0004654627
   alienation        allow       allows      almost        alone       along     alphabet      already        also
0.0002792776 0.0004654627 0.0002792776 0.0013963880 0.0004654627 0.0002792776 0.0002792776 0.0006516477 0.0013963880
     although     altogether     always     ambitions     american       among       amount     amygdalitis    analogous
0.0004654627 0.0004654627 0.0006516477 0.0002792776 0.0004654627 0.0012102029 0.0002792776 0.0002792776 0.0008378328
      analysis        analyze      ancients     anecdote      anguish       animal      another        answer      antiquity
0.0002792776 0.0004654627 0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0013963880 0.0006516477 0.0002792776
     anxieties       anything      apparatus    apparitions       appeals       appear      appeared      appearing       appears
0.0002792776 0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0010240179 0.0002792776 0.0002792776 0.0010240179
        apples    application     appropriate         arise        arisen          arm       around       arrange      artigues
0.0002792776 0.0002792776 0.0002792776 0.0004654627 0.0002792776 0.0002792776 0.0010240179 0.0004654627 0.0002792776
      artistic         aside          ask        asked        asleep     aspirations       aspire    assemblages      assembly
0.0002792776 0.0002792776 0.0008378328 0.0002792776 0.0013963880 0.0002792776 0.0004654627 0.0002792776 0.0004654627
```

2.resources_path

Get name of the resources location

```
> resources_path(local_path = getOption("corpustools_resources", NULL))
[1] "C:/Users/tangl/Documents/R/win-library/3.6/corpustools/ext_resources"
```

3.get_stopwords

To get a character vector of stopwords

```
> en_stop = get_stopwords('english')
> en_stop
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"        "your"       "yours"      "yourself"   "yourselves"
 [14] "he"         "him"        "his"        "himself"    "she"        "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"
 [27] "their"      "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"      "those"      "am"         "is"
 [40] "are"        "was"        "were"       "be"         "been"       "being"      "have"       "has"        "had"        "having"     "do"         "does"       "did"
 [53] "doing"      "would"      "should"     "could"      "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"      "they're"    "i've"
 [66] "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"      "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"
 [79] "we'll"      "they'll"    "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"     "won't"
 [92] "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"      "that's"     "who's"      "what's"     "here's"     "there's"
[105] "when's"     "where's"    "why's"      "how's"      "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"
[118] "until"      "while"      "of"         "at"         "by"         "for"        "with"       "about"      "against"    "between"    "into"       "through"    "during"
[131] "before"     "after"      "above"      "below"      "to"         "from"       "up"         "down"       "in"         "out"        "on"         "off"        "over"
[144] "under"      "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"        "how"        "all"        "any"
[157] "both"       "each"       "few"        "more"       "most"       "other"      "some"       "such"       "no"         "nor"        "not"        "only"       "own"
[170] "same"       "so"         "than"       "too"        "very"
```

(2)Stringi:

1.stri_enc_isascii

To check if a Data Stream Is Possibly in ASCII, the function checks whether all bytes in a string are <= 127.

```
> stri_enc_isascii(SAT1txt)
  [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [29]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [57]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [85]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[113]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[141] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[169]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[197]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[225]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[253]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[281]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[309]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[337]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[365]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[393]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[421]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[449]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[477]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[505]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[533]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[561]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[589]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[617]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[645]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[673]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[701]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[729]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[757]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[785]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[813]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[841]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[869]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

2.stri_remove_empty

To remove all empty strings from a character vector

```
> empSAT<-stri_remove_empty(SAT1txt, na_empty = FALSE)
> empSAT
 [1] "ï»¿"                                                                   "INTRODUCTION"
 [3] "Before the dawn of history mankind was engaged in the study of dreaming." "The wise man among the ancients was preÃ«minently the interpreter of"
 [5] "dreams. The ability to interpret successfully or plausibly was the"      "quickest road to royal favor, as Joseph and Daniel found it to be;"
 [7] "failure to give satisfaction in this respect led to banishment from"     "court or death. When a scholar laboriously translates a cuneiform tablet"
 [9] "dug up from a Babylonian mound where it has lain buried for five"        "thousand years or more, the chances are that it will turn out either an"
[11] "astrological treatise or a dream book. If the former, we look upon it"   "with some indulgence; if the latter with pure contempt. For we know that"
[13] "the study of the stars, though undertaken for selfish reasons and"       "pursued in the spirit of charlatanry, led at length to physical science,"
[15] "while the study of dreams has proved as unprofitable as the dreaming of" "them. Out of astrology grew astronomy. Out of oneiromancy has"
[17] "grown--nothing."                                                         "That at least was substantially true up to the beginning of the present"
[19] "century. Dream books in all languages continued to sell in cheap"        "editions and the interpreters of dreams made a decent or, at any rate, a"
[21] "comfortable living out of the poorer classes. But the psychologist"      "rarely paid attention to dreams except incidentally in his study of"
[23] "imagery, association and the speed of thought. But now a change has come" "over the spirit of the times. The subject of the significance of dreams,"
[25] "so long ignored, has suddenly become a matter of energetic study and of"  "fiery controversy the world over."
[27] "The cause of this revival of interest is the new point of view brought"  "forward by Professor Bergson in the paper which is here made accessible"
[29] "to the English-reading public. This is the idea that we can explore the" "unconscious substratum of our mentality, the storehouse of our memories,"
[31] "by means of dreams, for these memories are by no means inert, but have,"  "as it were, a life and purpose of their own, and strive to rise into"
[33] "consciousness whenever they get a chance, even into the"                 "semi-consciousness of a dream. To use Professor Bergson's striking"
[35] "metaphor, our memories are packed away under pressure like steam in a"   "boiler and the dream is their escape valve."
```

3.stri_escape_unicode

To escapes all Unicode (not ASCII-printable) code points.

```
> escapeSAT
  [1] "\\u00ef\\u00bb\\u00bf"
  [2] ""
  [3] "INTRODUCTION"
  [4] ""
  [5] ""
  [6] "Before the dawn of history mankind was engaged in the study of dreaming."
  [7] "The wise man among the ancients was pre\\u00c3\\u00abminently the interpreter of"
  [8] "dreams. The ability to interpret successfully or plausibly was the"
  [9] "quickest road to royal favor, as Joseph and Daniel found it to be;"
 [10] "failure to give satisfaction in this respect led to banishment from"
 [11] "court or death. When a scholar laboriously translates a cuneiform tablet"
 [12] "dug up from a Babylonian mound where it has lain buried for five"
 [13] "thousand years or more, the chances are that it will turn out either an"
 [14] "astrological treatise or a dream book. If the former, we look upon it"
 [15] "with some indulgence; if the latter with pure contempt. For we know that"
 [16] "the study of the stars, though undertaken for selfish reasons and"
 [17] "pursued in the spirit of charlatanry, led at length to physical science,"
 [18] "while the study of dreams has proved as unprofitable as the dreaming of"
 [19] "them. Out of astrology grew astronomy. Out of oneiromancy has"
 [20] "grown--nothing."
 [21] ""
 [22] "That at least was substantially true up to the beginning of the present"
 [23] "century. Dream books in all languages continued to sell in cheap"
 [24] "editions and the interpreters of dreams made a decent or, at any rate, a"
 [25] "comfortable living out of the poorer classes. But the psychologist"
 [26] "rarely paid attention to dreams except incidentally in his study of"
 [27] "imagery, association and the speed of thought. But now a change has come"
 [28] "over the spirit of the times. The subject of the significance of dreams,"
 [29] "so long ignored, has suddenly become a matter of energetic study and of"
 [30] "fiery controversy the world over."
 [31] ""
 [32] "The cause of this revival of interest is the new point of view brought"
 [33] "forward by Professor Bergson in the paper which is here made accessible"
 [34] "to the English-reading public. This is the idea that we can explore the"
 [35] "unconscious substratum of our mentality, the storehouse of our memories,"
 [36] "by means of dreams, for these memories are by no means inert, but have,"
 [37] "as it were, a life and purpose of their own, and strive to rise into"
 [38] "consciousness whenever they get a chance, even into the"
 [39] "semi-consciousness of a dream. To use Professor Bergson\\'s striking"
 [40] "metaphor, our memories are packed away under pressure like steam in a"
 [41] "boiler and the dream is their escape valve."
```

(3)Quanteda

1.sparsity

To compute the sparsity of a document-feature matrix, return the proportion of sparseness of a document-feature matrix, equal to the proportion of cells that have zero counts.

```
> #quanteda
> #Compute the sparsity of a document-feature matrix
> sparsity(SATdfm)
[1] 0.9947129
>
```

2.topfeatures

To list the most (or least) frequently occurring features in a dfm, either as a whole or separated by document.

```
> #Identify the most frequent features in a dfm
> topfeatures(SATdfm, n = 10, decreasing = TRUE, scheme = c("count","docfreq"), groups = NULL)
 the    ,    .   of   to  and   in    a   is that
 718  696  454  438  287  231  214  203  194  193
>
```

3.nsyllable

Returns a count of the number of syllables in texts. For English words, the syllable count is exact and looked up from the CMU pronunciation dictionary, from the default syllable dictionary data_int_syllables. For any word not in the dictionary, the syllable count is estimated by counting vowel clusters.

```
> #Count syllables in a text
> SAT1tokens[22]
tokens from 1 document.
text22 :
 [1] "That"          "at"           "least"        "was"          "substantially" "true"         "up"           "to"           "the"          "beginning"
[11] "of"           "the"          "present"

> nsyllable(SAT1tokens[22], syllable_dictionary = quanteda::data_int_syllables,use.names = FALSE)
$text22
 [1] 1 1 1 1 4 1 1 1 1 3 1 1 2
```

2.h.2 The theme of this article

This article is mainly about dreams. From part 2.c above, we can see that the word "dream" appears in almost every paragraph and has the highest frequency. In the introduction part, there is a lot of content about the understanding of dreams by different people of different eras. In the body part, there's a detailed explanation of how outside stimuli of different types can affect the dreaming process. The second most common word is “memory”, the author explores the relationship between dreams and memory. Obviously, memory has a major impact on dreams. Other words that appear frequently and are worth noting are "consciousness", "sensation", "see", "life". We can see that the author is trying to explain dreams and relate them to real-life experience.

3. Word Search

To obtain the document number, the line number and word index in the sentence, we write a function to work on it. The input of this function is the searching word/phase in the string and the searching path in the string.

After finishing all the preparations for the variable, we start searching the text.

```
for(i in 1:length(SAT)){
  tmp <-SAT[[i]]$content
  mycorpus <- VCorpus(VectorSource(tmp))
  corpus_frame <- data.frame(text=unlist(sapply(mycorpus, `[`, "content")), stringsAsFactors=F)
  sentences <-chunk_into_sentences(corpus_frame)
  sentences <- removePunctuation(sentences)
  sentences <- removeNumbers(sentences)

  for(j in 1:length(sentences)){
    tmp1 <-tokens(sentences)[j]
    tmp2 <- unlist(tmp1, use.names=FALSE)
    for(m in 1:length(tmp2)){
      count <-0
      n <-1
      while(n<=wordNo){
        if(wtmp[n] == toString(tmp2[m+n-1])){
          count = count+1
        }
        n=n+1
      }
      if(count == wordNo){
        sentInd <-append(sentInd,toString(m))
      }
    }
  }
```

The first for-loop is to find word index in the sentence. Firstly it traversal all the sentences and then obtain the word token from each of the sentences. Then it tries to pair the word or phrase with the tokens. If the searching word/phase is same as one of the tokens, (If it is a phase, it tries to pair both word and order), the word index (if it is a phrase, it will same the first-word index in the sentence) will be saved and output. This searching detects upper and lower case of word's characters.

```
    # Remove punctuation
    tmp <-removePunctuation(tmp)
    # Remove numbers
    tmp <-removeNumbers(tmp)
    for(j in 1:length(tmp)){
      tmp1 <-tokens(tmp)[j]
      tmp2 <- unlist(tmp1, use.names=FALSE)
      for(m in 1:length(tmp2)){
        count <-0
        n <-1
        while(n<=wordNo){
          if(wtmp[n] == toString(tmp2[m+n-1])){
            count = count+1
          }
          n=n+1
        }
        if(count == wordNo){
          documentNo <-append(documentNo,toString(i))
          lineNo <-append(lineNo,toString(j))
          wordInd <-append(wordInd,toString(m))
        }
      }
    }
    # Divide each document into sentences
    mycorpus <- VCorpus(VectorSource(tmp))
    corpus_frame <- data.frame(text=unlist(sapply(mycorpus, `[`, "content")), stringsAsFactors=F)
    sentences <-chunk_into_sentences(corpus_frame)
  }
  return(list(paste("Document Number = ", documentNo),
              paste("Line Number = ", lineNo),
              paste("Word Index in line = ", wordInd),
              paste("Word Index in sentence = ", sentInd)))
```

Same things we did for the document number and line number. But for these, we do not need to divide sentences.

To test our functions, we give three test cases (including words and phrases more than six characters), in two conditions. We use the entire text as our first condition. For we only have one entire text but we need to test for document number. Therefore, we separate the whole passage into three texts as three documents in a folder as our second condition.

3.1 Example for the word “efficacious”

```
# Testcase 1-1
Find_wordphase("efficacious","/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/")

# Testcase 1-2
Find_wordphase("efficacious","/Users/tianyuyang/Desktop/project2/test/")
```

```
[[1]]
[1] "Document Number =  1"

[[2]]
[1] "Line Number =  156"

[[3]]
[1] "Word Index in line =  7"

[[4]]
[1] "Word Index in sentence =  25"

The working directory was changed to /Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt inside a notebook
chunk. The working directory will be reset when the chunk is finished running. Use the knitr root.dir option in the
setup chunk to change the working directory for notebook chunks.[[1]]
[1] "Document Number =  1"

[[2]]
[1] "Line Number =  156"

[[3]]
[1] "Word Index in line =  7"

[[4]]
[1] "Word Index in sentence =  25"
```

In the first condition, only one entire text in the path. The search word is “efficacious” and the search path is "/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/".

As it is shown as in the figure above, the word “efficacious” is in the Document 1, Line Number is 156 and its line word index is 7. Its word index in its sentence is 25.

In the second condition, there are three divided documents in the path. The search word is “efficacious” and the search path is "/Users/tianyuyang/Desktop/project2/test/".

As it is shown as in the figure above, the word “efficacious” is in the Document 1, Line Number is 156 and its line word index is 7. Its word index in its sentence is 25.

3.2 Example for the word “materials”

```
# Testcase 2-1
Find_wordphase("materials","/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/")

# Testcase 2-2
Find_wordphase("materials","/Users/tianyuyang/Desktop/project2/test/")
```

```
[[1]]
[1] "Document Number =  1" "Document Number =  1" "Document Number =  1"

[[2]]
[1] "Line Number =  432" "Line Number =  432" "Line Number =  485"

[[3]]
[1] "Word Index in line =  1"  "Word Index in line =  10" "Word Index in line =  1"

[[4]]
[1] "Word Index in sentence =  5"  "Word Index in sentence =  6"  "Word Index in sentence =  24"

The working directory was changed to /Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt inside a notebook
chunk. The working directory will be reset when the chunk is finished running. Use the knitr root.dir option in the
setup chunk to change the working directory for notebook chunks.[[1]]
[1] "Document Number =  2" "Document Number =  2" "Document Number =  2"

[[2]]
[1] "Line Number =  58"  "Line Number =  58"  "Line Number =  111"

[[3]]
[1] "Word Index in line =  1"  "Word Index in line =  10" "Word Index in line =  1"

[[4]]
[1] "Word Index in sentence =  5"  "Word Index in sentence =  6"  "Word Index in sentence =  24"
```

In the first condition, only one entire text in the path. The search word is “materials” and the search path is "/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/".

As it is shown as in the figure above, there are totally three “materials”. The first word “materials” is in the Document 1, Line Number is 432 and its line word index is 1. Its word index in its sentence is 5. The second word “materials” is in the Document 1, Line Number is 432 and its line word index is 10. Its word index in its sentence is 6. The third word “materials” is in the Document 1, Line Number is 485 and its line word index is 1. Its word index in its sentence is 24.

In the second condition, there are three divided documents in the path. The search word is “materials” and the search path is "/Users/tianyuyang/Desktop/project2/test/".

As it is shown as in the figure above, there are totally three “materials”. The first word “materials” is in the Document 2, Line Number is 58 and its line word index is 1. Its word index in its sentence is 5. The second word “materials” is in the Document 2, Line Number is 58 and its line word index is 10. Its word index in its sentence is 6. The third word

“materials” is in the Document 2, Line Number is 111 and its line word index is 1. Its word index in its sentence is 24.

3.3 Example for the phrase “a host of”

```
# Testcase 3-1
Find_wordphase("a host of","/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/")

# Testcase 3-2
Find_wordphase("a host of","/Users/tianyuyang/Desktop/project2/test/")
```

```
[[1]]
[1] "Document Number =  1" "Document Number =  1"

[[2]]
[1] "Line Number =  424" "Line Number =  798"

[[3]]
[1] "Word Index in line =  4"  "Word Index in line =  11"

[[4]]
[1] "Word Index in sentence =  14" "Word Index in sentence =  4"

The working directory was changed to /Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt inside a notebook
chunk. The working directory will be reset when the chunk is finished running. Use the knitr root.dir option in the
setup chunk to change the working directory for notebook chunks.[[1]]
[1] "Document Number =  2" "Document Number =  3"

[[2]]
[1] "Line Number =  50"  "Line Number =  112"

[[3]]
[1] "Word Index in line =  4"  "Word Index in line =  11"

[[4]]
[1] "Word Index in sentence =  14" "Word Index in sentence =  4"
```

In the first condition, only one entire text in the path. The search word is “a host of” and the search path is "/Users/tianyuyang/Desktop/project2/CSCE6444_Project2/txt/".

As it is shown as in the figure above, there are totally two “a host of”. The first phrase “a host of” is in the Document 1, Line Number is 424 and its line word index is 4. Its word index in its sentence is 14. The second phrase “a host of” is in the Document 1, Line Number is 798 and its line word index is 11. Its word index in its sentence is 4.

In the second condition, there are three divided documents in the path. The search word is “a host of” and the search path is "/Users/tianyuyang/Desktop/project2/test/".

As it is shown as in the figure above, there are totally two “a host of”. The first phrase “a host of” is in the Document 2, Line Number is 50 and its line word index is 4. Its word index in its

sentence is 14. The second phrase “a host of” is in the Document 3, Line Number is 112 and its line word index is 11. Its word index in its sentence is 4.

4. Conclusion

In this project, we have done a lot of work on Natural language processing. We mainly learned and used several packages to process this book. By doing this project, we deepened our understanding of the R and Natural language processing, and we improved our R programming capabilities.

In part a, we tried all functions in lecture 4, and we had a basic understanding of this book.

In part b, we tried to find the 10 longest sentences in this book. First, we got the vectors contains every sentence in the text and then wrote two R functions to realize the function.

In part c. we displayed the dendrogram and the “WordCloud” for each paragraph. “WordCloud” is a very powerful package. We got two kinds of important graphs for text and they are very intuitive and contain a lot of information. From these graphs, we can easily get the frequency of words used in each paragraph. We realize that data visualization is both an art and a science that makes complex data easier to understand and use.

In part d, we found the longest word and longest sentence in each of the paragraphs. We have further enhanced our ability to process texts and have a better understanding of the structure of this article.

In part e, we used a package called "WordNet", it is a large lexical database of English. First, we got the words that contain no less than 5 characters and then we used this package's function to distinguish whether a word is a noun or a verb.

In part f, we used package “zipfR” to analyze word frequency. It has powerful statistical models and utilities for the analysis of word frequency distributions.

In part g, we generated bigrams and trigrams for all words in the first three paragraphs. To do that, we used a function from the package "quanteda". It's a very useful package to manage and analyze textual data.

In part h, we used “corpusTools”, “stringi”, and “quanteda”. We chose three functions from each package and apply them to this book. These packages provide tools that can give us a higher understanding of the text. Also, we tried to understand the theme of the book through the work we did before. This book is mainly about the author's explanations about

dreams. It's remarkably efficient to understand what a book is about by some simple applications of NLP.

Further, we have written some functions to search for a specific word or phrase in the documents. This is a very important and common feature that we use every day. Through the implementation of this function in R, our programming ability has been improved.

We believe that Natural language processing is a very important subject. Through it, we can efficiently process some long texts to get a general understanding of its content. Computers are great at working with standardized and structured data like database tables and financial records. But it’s not an easy task teaching machines to understand how we communicate. By using the technology of Natural Language Processing, we can aid computers to understand the human’s natural language.

In the future, we intend to learn more about NLP, especially in areas related to deep learning, because it is very interesting and powerful. Also, we realize that R is easily extensible through functions and extensions, and there are so many useful packages. But it's more important to choose the appropriate package and combine its functionality with your goals.